Disaggregation of population estimates at high spatial resolution with machine learning

Georgati, Marina

DOI (link to publication from Publisher):
10.54337/aau679682906

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Georgati, M. (2023). *Disaggregation of population estimates at high spatial resolution with machine learning*. Aalborg Universitetsforlag. https://doi.org/10.54337/aau679682906

DISAGGREGATION OF POPULATION ESTIMATES AT HIGH SPATIAL RESOLUTION WITH MACHINE LEARNING

BY
MARINA GEORGATI

DISSERTATION SUBMITTED 2023

AALBORG UNIVERSITY
DENMARK

Disaggregation of population estimates at high spatial resolution with machine learning

Ph.D. Dissertation

Marina Georgati

Dissertation submitted October 26, 2023

Dissertation submitted:	November 2023
PhD supervisor:	Prof. Carsten Keßler Aalborg University
Assistant PhD supervisor:	Prof. Henning Sten Hansen Aalborg University
PhD committee:	Professor Anja Jørgensen (chair) Aalborg University, Denmark Research Professor Marta Bivand Erdal Peace Research Institute Oslo, Norway Associate Professor Evelyn Uuemaa University of Tartu, Estonia
PhD Series:	Technical Faculty of IT and Design, Aalborg University
Department:	Department of Sustainability and Planning

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-598-3

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Stibo Complete, 2023

Curriculum Vitae

Marina Georgati

Marina received her M.Sc. in Architecture from the National Technical University of Athens, Greece, in 2017. In 2019, she received her M.Sc. in Geoinformatics from Aalborg University, Copenhagen, Denmark. In the same year, she joined the Department of Planning at Aalborg University as a research assistant. She started her PhD fellowship at Aalborg University in February 2020, working on cohort-based population estimations at the local level in European cities within the scope of the Horizon 2020 'Future Migration Scenarios for Europe' (FUME) project. Her research interests revolve around the use of programming for geospatial, urban and socio-economic analysis. She strongly believes in the potential of AI technologies to support urban development processes, making cities more sustainable and liveable for all. She enjoys experimenting with 3D visualisations, routing services and webGIS applications.

Abstract

Migration is not just about moving from one place to another; it involves the transformation of destination cities, where the amalgamation of cultures and ideas reshapes the urban landscape. In these vibrant spaces, diversity flourishes, as newcomers infuse fresh perspectives, enriching the social fabric and driving economic growth. Nevertheless, destination cities also deal with complexities associated with migration. Segregation and stigmatisation exacerbate, while issues such as exploitation of labour, rising real estate prices and gentrification pose significant threats for social cohesion.

As a result, crucial questions arise regarding the residential patterns of different demographic groups, especially migrants, within future cities and how cities can effectively manage migration to ensure a sustainable and inclusive future.

This PhD dissertation delves into this complex issue as part of a broader interdisciplinary project on '*Future Migration Scenarios for Europe*' *(FUME).* The study aims to contribute with the development of methods and tools for estimating migrant settlements at local level and explaining the underlying determinants of their distributions. These methods can enhance population forecasting processes and provide insights for better decision-making and more inclusive urban planning. By breaking down the challenge into a spatial disaggregation problem, the dissertation applies machine learning techniques to improve the accuracy of estimations at high spatial granularity. Furthermore, it aims to contribute to the exploration of potential determining factors behind these settlement patterns by using interpretation measures through machine learning. The research also seeks to shed light on patterns and dynamics of migration by leveraging geospatial visualisation techniques.

The main objective of this project is to explore methods for downscaling population counts of various cohorts from coarse data to fine-grained cells and interpret the significance of underlying socio-demographic and topological patterns on the distribution of these cohorts, especially of migrants.

The dissertation is structured around three research questions, each of which serves as a cornerstone for the contributions made in this study. First, it demonstrates how machine learning can be an efficient tool for simultaneously downscaling multiple population cohorts. The results show that the proposed approach is convenient and efficient, outperforming traditional disaggregation approaches (e.g. pychnophylactic interpolation). The developed model breaks the smooth distributions produced by traditional approaches and improves the spatial variability of the outputs by capturing complex relationships and dependencies between the target variables. The thesis also advances over previous research into the spatial evaluation of the disaggregated estimates. In addition to common evaluation metrics, it uses the concept of individualised neighbourhoods to interpret the spatial errors and enhance the predictive accuracy and generalisability of spatial models. It then applies different visualisation and machine learning techniques to explore residential patterns and relationships among various demographic groups.

This dissertation showcases methodological advances in the field of spatial disaggregation and modelling. By building upon existing tools and achieving fine-grained estimates of migrant populations for destination cities in Europe, the dissertation goes beyond the capabilities of existing methods. It also suggests methods for the analysis of rich datasets and provides insights into the interpretation of machine learning models. This type of research is valuable, as it enhances the understanding and modelling of complex spatial phenomena, in this case, the distribution of migrant populations at a detailed level within destination cities. The fine-grained estimates obtained through the developed tool can offer significant advantages for spatial analysis, enabling policy makers and researchers to make more informed decisions and design more targeted interventions.

This dissertation provides a starting point for more efficient and accurate high-resolution estimates of historical and future population distributions. Finally, it recommends alternative paths for future research on refining the proposed methodology, but also on population projections with time series of rasters and on how collaboration between academia and local authorities can be promoted to engage local authorities in the production of more relevant and applicable population estimates in urban planning.

Resumé

Migration handler ikke blot om at flytte fra et sted til et andet, men omhandler også transformation af destinationsbyerne, hvor sammensmeltningen af kulturer og ideer omformer det urbane landskab. Mangfoldigheden blomstrer i disse pulserende områder, når tilflytterne tilfører nye perspektiver, beriger den sociale struktur og fremmer økonomiske vækst. Ikke desto mindre står destinationsbyerne også over for kompleksiteter forbundet med migration. Segregering og stigmatisering aggraveres, mens aspekter såsom udnyttelse af arbejdskraft, stigende ejendomspriser og gentrificering skaber betydelige udfordringer for den sociale samhørighed.

Som følge heraf opstår afgørende spørgsmål om boligmønstrene for forskellige demografiske grupper, især migranter, i fremtidens byer og om, hvordan byer effektivt kan håndtere migration for at sikre en bæredygtig og inkluderende fremtid.

Denne ph.d.-afhandling dykker ned i dette komplekse emne som en del af et bredere tværfagligt projekt om '*Future Migration Scenarios for Europe*' *(FUME)*. Undersøgelsen har til formål at bidrage til udviklingen af metoder og værktøjer til estimering af migrantbosættelser på lokalt niveau og forklare de underliggende determinanter for deres fordeling. Metoderne kan forbedre befolkningsprognoser og give indsigt, til at forbedre beslutningstagning og lave mere inkluderende byplanlægning. Ved at nedbryde undersøgelsen til et rumligt disaggregeringsproblem, anvendes maskinlæringsteknikker for at forbedre nøjagtigheden af estimater ved høj rumlig granularitet. Endvidere har afhandlingen til formål at bidrage til undersøgelsen af potentielle afgørende faktorer, som danner baggrund for bosætningsmønstrene ved at anvende fortolkningsmetoder gennem maskinlæring. Forskningen forsøger også at belyse migrationsmønstre og -dynamikker via anvendelse af geospatiale visualiseringsteknikker. Hovedformålet med dette projekt er at undersøge metoder til disaggregering af befolkningstal for forskellige grupper fra grovkornede data til finkornede celler samt fortolke betydningen af underliggende sociodemografiske og topologiske mønstre som karakteriserer fordelingen af disse kohorter, med et specifikt fokus på migranter.

Afhandlingen er struktureret omkring tre forskningsspørgsmål, som hver især fungerer som hjørnesten i forbindelse med undersøgelsen. For det første demonstrerer den, hvordan maskinlæring kan være et effektivt værktøj til simultant disaggregering af flere befolkningskohorter. Resultaterne viser, at den foreslåede tilgang er praktisk og effektiv og overgår traditionelle disaggregeringsmetoder (f.eks. pychnophylactic interpolation). Den udviklede model bryder de glatte fordelinger, der produceres af traditionelle metoder, og forbedrer den rumlige variabilitet af resultaterne ved at indfange komplekse forhold og afhængigheder mellem målvariablerne. Afhandlingen gør også fremskridt i forhold til tidligere forskning i den rumlige evaluering af de disaggregerede estimater. Udover almindelige evalueringsmålinger bruger den begrebet individualiserede nabolag til at fortolke de rumlige fejl og forbedre den prædiktive nøjagtighed og generaliserbarhed af rumlige modeller. Forskellige visualiserings- og maskinlæringsteknikker anvendes til udforskning af bosætningsmønstre og relationer mellem forskellige demografiske grupper.

Denne afhandling fremviser metodologiske fremskridt inden for rumlig disaggregering og modellering. Ved at bygge på eksisterende værktøjer og opnå detaljerede estimater af migrantbefolkninger for destinationsbyer i Europa, går afhandlingen ud over de eksisterende metoder. Den foreslår ligeledes metoder til analyse af omfattende datasæt og giver indsigt i fortolkningen af maskinlæringsmodeller. Denne type forskning er værdifuld, da den forbedrer forståelsen og modelleringen af komplekse rumlige fænomener, i dette tilfælde fordelingen af migrantbefolkninger, på et detaljeret niveau i destinationsbyer. De detaljerede estimater, der opnås gennem det udviklede værktøj, kan give betydelige fordele i forbindelse med rumlige analyser. Dette muliggør, at beslutningstagere og forskere kan træffe mere velinformerede beslutninger og designe mere målrettede indgreb.

Denne afhandling giver et udgangspunkt for mere effektive og præcise estimater i høj opløsning af historiske og fremtidige befolkningsfordelinger. Endelig anbefaler den alternative veje til fremtidig forskning i at forfine den foreslåede metode, men også i befolkningsprognoser med tidsserier af raster og i, hvordan samarbejde mellem den akademiske verden og lokale myndigheder kan fremmes, for at engagere lokale myndigheder i udarbejdelsen af mere relevante og anvendelige befolkningsestimater i byplanlægning.

Contents

Abbreviations

ANNs Artificial neural networks.

CNN Convolutional neural networks.

CoB Country of birth.

CoC Country of citizenship.

CoDa Compositional data.

CoO Country of origin.

CPR Central Person Register.

DL Deep learning.

DNNs Deep neural networks.

DST Statistics Denmark.

EU European Union.

FUME Future Migration Scenarios for Europe.

GAN Generative adversarial networks.

GB Gradient boosting.

GDP Gross domestic product.

GeoAI Geographic artificial intelligence.

GIS Geographic Information System.

GRUMP Global Rural-Urban Mapping Project.

IIASA International Institute for Applied Systems Analysis.

LSTM Long short-term memory.

MAE Mean absolute error.

MAPE Mean absolute percentage error.

MAUP Modifiable aerial unit problem.

MENA Middle East and North Africa.

ML Machine learning.

MSE Mean squared error.

NIDI Netherlands Interdisciplinary Demographic Institute.

PIK Potsdam Institute for Climate Impact Research.

RCPs Representative Concentration Pathways.

REE Relative error estimation.

RF Random forest.

RMSE Root mean square error.

RNN Recurrent neural networks.

SRES Special Report on Emissions Scenarios.

SSPs Shared Socioeconomic Pathways.

UEK Kraków University of Economics.

UK United Kingdom.

UN United Nations.

UNIMAN University of Manchester.

UNSD United Nations Statistics Division.

US United States.

VIM Variable importance measure.

WIC Wittgenstein Centre.

XAI Explainable artificial intelligence.

Preface

Five years ago, the idea of pursuing a PhD would have elicited an absolute "No way!" from me. However, life is unpredictable and led me to a remarkable opportunity that I could not pass up. Aalborg University in Copenhagen presented an exceptional chance that intrigued me deeply, combining computer sciences with urban and socio-economic analysis – an area of interest that resonated with my background in architecture and a fresh Master's degree in Geographic Information System (GIS).

Motivated by the significance of informed decision-making in urban planning and predictive modelling (in an unpredictable world), I embarked on this PhD project to bridge these fields. My aim is to leverage GIS capabilities for data-driven analysis to promote efficient, sustainable and inclusive urban environments. As a migrant myself, I have personally experienced the profound impact of well-planned cities on individuals and communities. This research project reflects my personal connection to the topic and my determination to contribute to the transformation of decision-making for urban development.

Filled with ambition, curiosity and boundless energy, I eagerly embraced this new chapter in my academic and personal growth. However, fate (once again) had its own plans, and as I stepped into this exciting position, the world was struck by the unprecedented COVID-19 pandemic. Suddenly my expectations were met with the harsh reality of restrictions and the challenges of working remotely.

Amidst these challenges, the journey has been both difficult and rewarding, as I have engaged in intriguing online and in-person discussions and research, and achieved meaningful results with the support of many individuals. By examining patterns, trends and correlations within vast datasets, I aimed to provide a more comprehensive understanding of population dynamics in urban environments.

In this dissertation, I will share the journey of how my passion, experience and motivation converged into a research project that combines population and computer sciences, machine learning and GIS analysis. I hope that my work will inspire future researchers and practitioners to embrace interdisciplinary approaches and pave the way for more informed and fair cities.

Information on funding

This PhD research has been jointly financed by Aalborg University and the FUME project (*Future Migration Scenarios for Europe*, *https:// futuremigration.eu*). FUME is funded by the EU (Grant number: 8706499).

Acknowledgements

This PhD journey has been a rollercoaster of emotions and experiences. There have been moments when it has felt stagnant, trapped within the walls of a room, while at other times it has raced forward, leaving little time for reflection. Yet, despite the challenges, it has been educational and inspiring. Throughout this journey of academic and personal development, I am sincerely grateful for the immense support of my academic mentors and personal advocates. Their constant encouragement and guidance have been my compass in this path of knowledge and personal growth and have pushed me to the finish line.

It has been a great opportunity to be part of the FUME project, which has kept me motivated and broadened my research horizons into different areas. I would like to sincerely thank all the FUME partners for the inspiring and interesting discussions and their commitment. It has been a great pleasure to work in this interdisciplinary project, to learn from many demography and migration researchers, to have memorable project meetings and to hear encouraging words.

I would like to thank Bruno Martins and João Monteiro for hosting me in Lisbon and guiding me through the details of the research. Thank you, Bruno, for your insightful feedback at the beginning of this work and for the many discussions and insights into machine learning.

I am especially grateful to my main supervisor, Carsten Keßler, for giving me the opportunity to do this PhD and for supporting me all these years. You provided me with invaluable feedback and guidance on my research steps, as well as psychological support when I was mostly discouraged and disappointed. Thank you for the thorough proofreading on the dissertation and papers. Special thanks to my co-supervisor Henning Sten Hansen for your valuable guidance, targeted advice and the wisdom you have shared throughout this journey. Thanks to the GEO research group and our close colleagues in the LAND group for their daily input and emotional support. Thanks to Åse and Kasper for proofreading the Danish abstract.

Finally, a big thank you to my family. To my parents, Voula and Manolis, thank you for your endless support and determination to see me succeed. To my brother, Giannis, thank you for your boundless energy and passion for research, which has been so inspiring. To my dearest friends, Maria and Giorgos, for standing by me during the most challenging moments of this journey and for sharing my enthusiasm for GIS and coding, and to Imke, for sharing the frustrations of a PhD. To my partner, Sæbo, for your love, patience and belief in me, which kept me going through difficult times, and for your help with proofreading.

Marina Georgati
Aalborg University, October 26, 2023

Part I

Introduction

Dissertation details

Papers included in this dissertation

- **Georgati, M.**, Monteiro, J., Martins, B., & Keßler, C. (2022). Spatial disaggregation of population subgroups leveraging self-trained Mmulti-output gradient boosting regression Trees. In 25th AGILE Conference on Geographic Information Science: Artificial Intelligence in the service of Geospatial Technologies (Vol. 3, pp. 1-14). Copernicus GmbH. AGILE GIScience. DOI: https://doi.org/10.5194/agile-giss-3-5-2022

- **Georgati, M.**, Monteiro, J., Martins, B., Keßler, C., & Hansen, H. S. Modelling population distribution: a visual and quantitative analysis of gradient boosting and deep learning models for multi-output spatial disaggregation. Under review - Transactions in GIS.

- **Georgati, M.**, Hansen, H. S., & Keßler, C. Random forest variable importance measures for spatial dynamics: case studies from urban demography. Under review - International Journal of Geo-Information.

- Kveladze, I., **Georgati, M.**, Keßler, C., & Hansen, H. S. (2023). Analytics of historical human migration patterns: use cases of Amsterdam and Copenhagen. Journal of Location Based Services. DOI: https://doi.org/10.1080/17489725.2023.2238658

Other contributions

- **Georgati, M.** (2020). Data archive with data prepared for the model for each case study. FUME Project Deliverable 5.1.

- **Georgati, M.**, & Keßler, C. (2021). Spatially explicit population projections: the case of Copenhagen, Denmark. In P. Partsinevelos, P. Kyriakidis, M. Kavouras (Eds.), 24th AGILE Conference on Geographic Information Science Copernicus Publications. AGILE GIScience Vol. 2.

- Stonawski, M., **Georgati, M.**, Crisci M., Wissen L., Brzozowski J., Keßler C. (2021). Structured dataset with geocoded data on location of immigrants and their characteristics and on pull factors of immigration and immigrant distribution in cities. FUME Project Deliverable 6.3.

- Elío, J., **Georgati, M.**, Hansen, H. S., & Keßler, C. (2022). Migration studies with a compositional data approach: a case study of population structure in the capital region of Denmark. In O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, C. Garau (Eds.), Computational Science and Its Applications - ICCSA 2022 Workshops, Proceedings: Malaga, Spain, July 4–7, 2022, Proceedings, Part III (pp. 576-593). Springer. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) Vol. 13379 LNCS.

- Stonawski, M., Brzozowski, J., Pędziwiatr, K., & **Georgati, M.** (2022). Investigating neighbourhood concentration of immigrants in Poland: explorative evidence from Kraków. Bulletin of Geography. Socio-Economic Series, June(56), 143-159.

- **Georgati, M.** (2023). Report on modelling approach and implementation. FUME Project Deliverable 5.4.

- Stonawski, M., **Georgati, M.**, Crisci, M., Wissen, L., & Brzozowski J. (2023). Report on migration and residential segregation in European cities – now and future – analysis and projections of immigration and spatial distribution of immigrants. FUME Project Deliverable 6.4.

- **Georgati, M.**, Stonawski, M. Keßler, C., & Hansen, H. S., A Comparative Study of Spatio-Temporal Segregation Patterns with Gridded Population Data in Denmark, the Netherlands, and Poland. - Working paper.

Conference participation

- **Georgati, M.**, & Keßler, C. Spatially explicit population projections: the case of Copenhagen, Denmark. 24th AGILE Virtual Conference, June 8 – 11, 2021.

- **Georgati, M.**, & Keßler, C. Spatially explicit population projections, 18th IMISCOE Annual Conference (Online), July 7 - 9, 2021.

- **Georgati, M.**, Monteiro, J., Martins, B., & Keßler, C. Spatial disaggregation of population subgroups leveraging self-trained multi-output gradient boosting regression trees, 25th AGILE Conference, June 14 - 17, 2022, Vilnius.

- **Georgati, M.**, Elío, J., Kveladze, I., Keßler, C., & Hansen, H. S. Analysing spatio-temporal migration structures with gridded population data in Copenhagen, Denmark, RGS-IBM Annual International Conference, Aug. 30 - Sept. 2, 2022, Newcastle.

- FUME Joined final conference, April 26 - 27, 2023, Brussels (VUB/EU Parliament).

1

Introduction

Introduction

Global migration has increased from approximately 153 million in 1990 [147] to around 281 million in 2020 [65]. Despite the impact of COVID-19, which resulted in 2 million fewer international migrants in 2020 than would otherwise have been expected, migration continues to play a substantial role in the development and transformation of cities. It drives urbanisation processes and shapes the image and growth of urban areas. Migrants contribute to cities' social and cultural diversity, increase their attractiveness for companies as they join the labour force [41, 140], expand their amenity values [41] and strengthen the international network between them [127, 109]. Consequently, migration has substantial social and economic impacts on cities.

Nonetheless, migration raises complexities associated to discrimination, stigma and segregation, which threaten social cohesion. Additionally, it drives developments such as rising real estate prices and gentrification [157]. It is concerning that even in countries that actively promote social equality and integration, such as Denmark, there are political and social discussions about '*ghettos*' and '*parallel societies*'. The Danish '*Ghetto Package*', which specifically targets 'non-western' minorities, jeopardises social cohesion and has initiated gentrification processes [12, 145, 110]. Paradoxically, these debates and actions emerge while migrants actively contribute to the country's economic development and population growth counterbalancing population ageing and low birth rates.

Over the next decades, population growth in high-income countries will be solely driven by migration [149, 17]. As their urban centres will become more dense [156, 152], important questions arise around ***where people, particularly migrants, will live within cities in the future*** and how cities can effectively manage migration for a sustainable and inclusive future.

Although there are currently no established methods that use existing knowledge about the factors that influence migrant settlement to help cities predict potential future residential locations of migrant groups, emerging technological advances such as machine learning (ML) hold great potential to address these issues. Integrating demographic research in geospatial and computational studies, with a specific focus on migration and population estimates, can have a significant impact on local communities and economies. Cities can develop models and tools that provide insights for decision-making and guidelines for inclusive urban development. Such interdisciplinary approaches and models include the identification of optimal locations for the settlement of migrants, preventing segregation and ghettoisation, while considering the safety and attractiveness of neighbourhoods. This can add great value especially for the integration of new influxes of migrants or refugees (e.g. Ukrainians moving to other European countries).

Challenges such as those outlined above can be addressed by turning national and regional population projections into local migrant distributions in urban areas. Local population estimates reveal to be a useful tool for urban planning and for managing social and cultural diversity. Ultimately, the use of advanced computational techniques can help cities create sustainable and inclusive environments for both migrants and native populations. By leveraging the power of data and technology, cities can make informed decisions that lead to more efficient resource allocation, improved social cohesion and enhanced economic growth.

Numerous institutions have worked on estimating future population at national and regional levels [39, 148, 86, 72, 40, 73]. However, knowledge of the actual spatial distribution of different demographic groups at the local level of the cities and neighbourhoods is still lacking. Enriching this spatial knowledge with demographic characteristics, especially those related to migration, offers great potential for a better understanding of urbanisation processes in the context of social and spatial demography.

This study helps to bridge this gap between national and regional population projections, local planning and the factors influencing migrant distributions. The study is based on the latest advancements in ML and contributes to the field of geographic artificial intelligence [GeoAI; 68] by extending available models to socio-demographic applications.

It is presented in an article-based dissertation consisting of an introduction in Part I, which sets the context in which the study was developed, and the core papers in Part II. The first two papers investigate methods for developing and assessing data-driven models for disaggregating population cohorts into fine-grained grids. The third paper contributes to the field of explainable artificial intelligence [XAI; 80, 20] by explaining and confirming the decisions of the ML model on the basis of existing knowledge. The last publication promotes various geovisualisation techniques for the analysis of human migration patterns.

The study was developed in the context of the '*Future Migration Scenarios for Europe*' (FUME) Horizon 2020 project, which is further described in the following subsection. Section 1.2 outlines the requirements specified by FUME and places the choice of methodology in the temporal context of the period. It also outlines the motivations within which the scope of the project was framed. The research objectives follow in section 1.3, along with the research questions that guide this dissertation. Section 1.4 provides the structure of the dissertation.

1.1 The PhD study in the context of the FUME project

The present study was part of the *FUME* Horizon 2020 project. The main objective of FUME was to address the knowledge gap in the literature on cross-scale determinants, which leads to oversimplified migration scenarios. The major goals of the project were, first, to identify the push and pull factors that explain migration patterns and, second, to explore how future migrant movements may potentially shape destination cities.

The project therefore conducted qualitative research in countries of origin (i.e. Ukraine, Senegal, Iraq, Tunisia) to understand the drivers and trajectories of migration. This part of the project highlighted the role of migration drivers, their interrelationships, but also their dependence on cultural, geographical and historical differences between regions [48]. The author was not involved in this qualitative exploration and the drivers of migration from countries of origin (CoOs) are beyond the scope of this dissertation. They are therefore not discussed further. However, further information can be found in Sobczak-Szelc *et al.* [135], Diop *et al.* [33], degli Uberti *et al.* [32] and Soboleva *et al.* [136].

Another research pillar of FUME followed a multi-level approach, quantifying potential international flows at the global level and focusing on the resulting population stocks at the national, regional and local levels in four selected case studies: Amsterdam – a Western European city with a centuries-long history of migration, Copenhagen – a Northern European capital with large Middle Eastern migrant communities, Kraków – an Eastern European city with a recent influx of Eastern European migrants from non-European Union (EU) countries, Rome – a Southern European capital with large Eastern European and Asian migrant communities. These cases are presented in more detail in section 3.

This quantitative phase of the project involved many different partners, including the International Institute for Applied Systems Analysis (IIASA), the Netherlands Interdisciplinary Demographic Institute (NIDI), Statistics Denmark (DST) and the Potsdam Institute for Climate Impact Research (PIK), and was developed sequentially, with the results of each modelling stage feeding into the next. Figure 1.1 illustrates the multi-dimensional approach of FUME, with more details on the methodological steps following in section 2. The right column of the figure shows the products of each modelling approach, which are then used as inputs into the subsequent model. The contributions of this PhD study are highlighted in blue.

Initially, a set of narratives and scenarios were developed based on the evaluation of existing scenarios (IIASA, NIDI), drawing on the findings of the qualitative study in the origin case studies and a Delphi survey – Kraków University of Economics (UEK), University of Manchester (UNIMAN), IIASA. A migration model was developed by PIK to project international flows based on the scenarios and assumptions on the potential future levels of national gross domestic product (GDP). These projections were incorporated into a multi-dimensional demographic cohort-component population projections model to estimate population stocks by country of birth (CoB), sex, age and education at the national level.

The national projections were downscaled to the regional level in a statistical demographic model [153], which also included assumptions about internal migration. This output was then used in the local ML-based model, which disaggregated the population cohorts into gridded distributions – the methodology for generating these disaggregated counts was the main scope of the present PhD project. Finally, the disaggregated counts, together with the collected historical data, were used to analyse historical and projected patterns of segregation in the case studies.

The interdisciplinary nature of FUME, combining demographic, geospatial and computational approaches, is crucial to gain comprehensive insights into the dynamics of migration and its impact on cities. By fostering collaboration between researchers and experts from different fields, FUME has enhanced the understanding of migration patterns in cities and countries of origin and destination, and developed innovative strategies to address the challenges and opportunities associated with migration in urban areas. More information on each of the modelling steps of FUME is discussed in section 2.

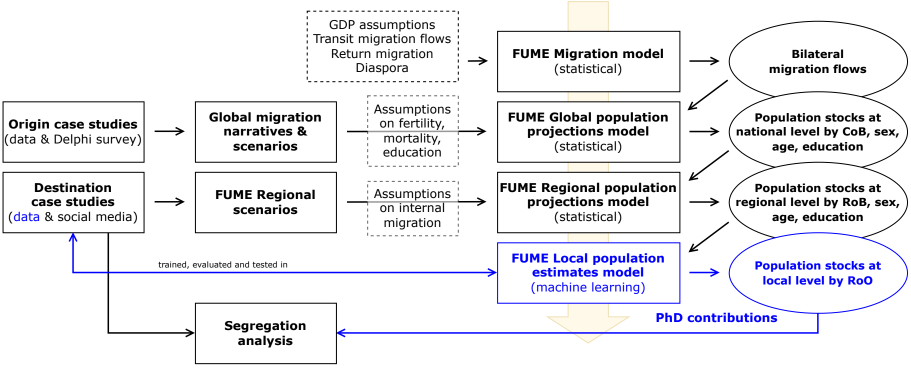


Fig. 1.1: Outline of the FUME approach and the contribution of this PhD study. The upper part involves the initial data collection in origin case studies for the global migration narratives, which inform the assumptions for the population projection model and national estimates. The lower part shows the data collection in destination case studies for the regional and local models. The right column displays the model outputs as inputs to the next model. The PhD contributions in blue focus on local modelling using machine learning and local-level population estimates by RoO. Abbreviations: GDP - gross domestic product, CoO - country of origin, RoO - region of origin.

1.2 Motivation and strategic choices

This subsection describes the requirements of FUME in the context of this PhD study and sets the methodological choices into perspective. The main objective of the spatial modelling within FUME was to develop a flexible ML-based methodological approach to project the future distribution of migrants at the local level of cities on a fine-grained grid. The methodology was to be applied to four case studies with special urban and multicultural characteristics and varying levels of data availability. ML modelling was chosen for its potential to recognise patterns between various variables such as the distribution of migrants, population size and various urban variables (e.g. land use, housing market data, development plans, income and education levels, unemployment, ethnic composition, religion and nearby services such as schools). The projections should be represented at a high spatial resolution of 100 metre (m) grid cells in multiple time steps.

Its ability to be connected to regional projections was one of the major requirements for the desired model. Additionally, it had to enable the estimation of multiple demographic dimensions while allowing for adjustments among the case studies and their corresponding demographic dimensions.

The initial idea evolved around training a model using historical data on the distribution of migrants at the same resolution and by individual CoO, such as a time series model. However, this approach faced challenges due to the uncertainty of collecting the required data during the COVID-19 pandemic. Even after the pandemic, the collection of these datasets was problematic as their availability and quality were questionable for some case studies. Additionally, there were doubts as to whether the available data would provide sufficient input to train a time series model such as a long short-term memory (LSTM) model, as described in section 2.5.

Additionally, during the initial stages of the study, there was a certain degree of ambiguity regarding the final projection output, mainly due to the conflation of migrant stocks and flows, as well as the challenges inherent in the migration modelling process. For this reason, an extensive literature review was deemed necessary to clarify concepts and examine methodologies. As someone without extensive knowledge of demographic theory and ML development, my initial step was to gain a clear understanding of the distinctions between population and migration projections, flows and stocks, while at the same time diving into ML concepts and programming. This review is presented in detail in section 2, with the concepts explored in each subsection leading to the approaches adopted in FUME for each step.

This narrative, like FUME, unfolds multi-dimensionally, from smaller to larger geographical scales. In section 2.1, it explores the significance of global and national population projections and their main modelling components (i.e. fertility, mortality, immigration and emigration; 30), with migration modelling playing a crucial role in population projections (sec. 2.2). As migration flows are much less predictable than mortality and fertility, many challenges arise. Focusing on the sub-national level, section 2.3 discusses cases and additional parameters, such as internal mobility, that can be incorporated into statistical demographic modelling. This, however, increases the complexity of the model and poses challenges in terms of data requirements at the examined spatial scale and the assumptions about the investigated factors.

As the scope of this PhD study was set at the local level, at 100 m resolution, section 2.4 focuses on population projections at finer spatial resolutions and discusses existing examples that are based either on gravity modelling with time-consuming processes or on geosimulations, that require simplifications of the analysis to aggregated population cohorts. Overall, existing conventional practices trade-off between the need for data that are not always easy to obtain, long processes, or the loss of spatial and contextual accuracy.

Given these limitations, this section concludes that ML and GeoAI can overcome such challenges and offer advantages for fine-grained population estimations. It demonstrates their potential to capture the complexities and interrelationships in the distributions of various cohorts through previous examples that have used remote sensing or disaggregation approaches to estimate population density.

Based on these observations, that are explicitly explained in the next section, the present work adopted an alternative perspective to the time series modelling that emphasised the simultaneous disaggregation of multiple population counts. This approach aimed to highlight the relationships and interactions between different population cohorts, and to leverage the available training data to inform the estimation process. It extended work previously conducted by Monteiro *et al.* [97, 98] and is described in detail in Papers A and B (sec. 8, 9).

Finally, it should also be mentioned that FUME's local partners were continuously collecting extensive population and building datasets, which were considered necessary inputs for the training. These datasets, often characterised by their heterogeneity, include a wide range of demographic, geographic and socio-economic variables. Papers C and D (sec. 10, 11) therefore focus on research approaches for studying migrant distribution patterns. These explorations offer advantages for gaining knowledge and enriching our understanding of the factors that influence the distribution of various cohorts. Exploring these rich datasets, identifying meaningful patterns and gaining important insights serve as a valuable resource for data-driven decision-making and policy formulation.

1.3 Research objectives

The aim of this dissertation is to investigate a ML-based method for fine-grained spatial disaggregation of different demographic groups within urban areas. It also aims to analyse the factors that influence the spatial distribution of different demographic groups within urban environments.

The main research objectives of this dissertation are to **explore innovative methods for generating and assessing disaggregated population estimates, and to examine patterns in the distribution of different migrant communities**.

The first objective is achieved through the development of a simulation tool based on self-training ML models that link aggregated population data to the specific residential locations of migrants in destination cities, while various assessment methods are examined for assessing the quality of the disaggregated estimates. In addition, ML and geovisualisation techniques are used to reveal patterns in the distribution of different migrant groups. The study primarily focuses on the cities of Amsterdam and Copenhagen. The dissertation is structured around three research questions in line with the above research objectives.

RQ1: How can multi-output ML models improve the disaggregation of multiple population variables over classic dasymetric approaches and single-output models?

RQ2: How can different assessment methods assist the spatial evaluation of disaggregated estimates when ground truth data are available at the target resolution?

RQ3: How can geovisualisation and ML techniques help to reveal patterns in the distribution of different migrant groups in relation to the native population and other socio-economic and urban characteristics?

The research questions are answered by the papers written during the PhD study. The full papers are included in Part II with their corresponding summaries in sections 4.1 - 4.4. Additional co-authored papers and reports were produced during the course of this study and supplement the dissertation by providing additional material for the analysis of the case study areas in section 3. The next subsection describes the structure of the dissertation and the connections between its different parts.

1.4 Structure

The dissertation consists of two parts: the introduction and the papers, where the four papers of the dissertation are presented (see Part II). Figure 1.2 shows the structure of the introduction (Part I) on the left, the research design in the middle and the contributions to the papers and reports developed during the project on the right. The introduction has seven sections, with section 1 introducing the context in which the research was developed and outlining the research objectives and questions. Section 2 presents the background to main concepts about migration and population projections, gridded population estimates and ML development. Section 3 presents the selected case studies, followed by section 4 containing the summaries of the papers. Section 5 briefly presents results not included in the papers on the estimated distribution of migrants based on selected FUME scenarios. Section 6 discusses the papers in relation to the research questions, contributions, limitations and future research. Section 7 concludes the first part of the dissertation.

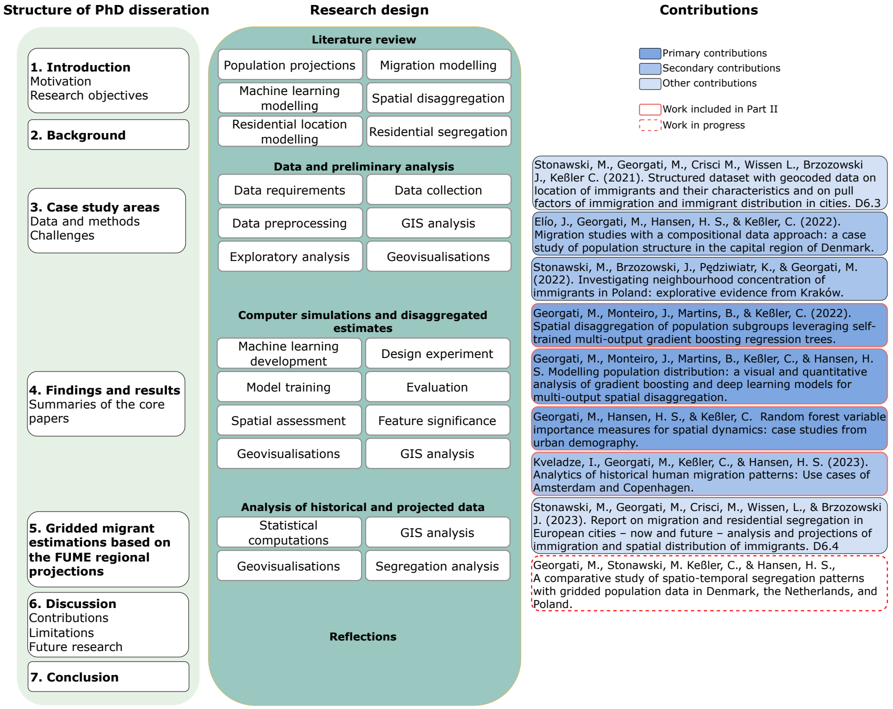

Structure of PhD disseration
1. Introduction
Motivation
Research objectives
2. Background
3. Case study areas
Data and methods
Challenges
4. Findings and results
Summaries of the core papers
5. Gridded migrant estimations based on the FUME regional projections
6. Discussion
Contributions
Limitations
Future research
7. Conclusion
Research design
Literature review
Population projections
Migration modelling
Machine learning modelling
Spatial disaggregation
Residential location modelling
Residential segregation
Data and preliminary analysis
Data requirements
Data collection
Data preprocessing
GIS analysis
Exploratory analysis
Geovisualisations
Computer simulations and disaggregated estimates
Machine learning development
Design experiment
Model training
Evaluation
Spatial assessment
Feature significance
Geovisualisations
GIS analysis
Analysis of historical and projected data
Statistical computations
GIS analysis
Geovisualisations
Segregation analysis
Reflections
Contributions
Primary contributions
Secondary contributions
Other contributions
Work included in Part II
Work in progress
Stonawski, M., Georgati, M., Crisci M., Wissen L., Brzozowski J., Keßler C. (2021). Structured dataset with geocoded data on location of immigrants and their characteristics and on pull factors of immigration and immigrant distribution in cities. D6.3
Elío, J., Georgati, M., Hansen, H. S., & Keßler, C. (2022). Migration studies with a compositional data approach: a case study of population structure in the capital region of Denmark.
Stonawski, M., Brzozowski, J., Pędziwiatr, K., & Georgati, M. (2022). Investigating neighbourhood concentration of immigrants in Poland: explorative evidence from Kraków.
Georgati, M., Monteiro, J., Martins, B., & Keßler, C. (2022). Spatial disaggregation of population subgroups leveraging self-trained multi-output gradient boosting regression trees.
Georgati, M., Monteiro, J., Martins, B., Keßler, C., & Hansen, H. S. Modelling population distribution: a visual and quantitative analysis of gradient boosting and deep learning models for multi-output spatial disaggregation.
Georgati, M., Hansen, H. S., & Keßler, C. Random forest variable importance measures for spatial dynamics: case studies from urban demography.
Kveladze, I., Georgati, M., Keßler, C., & Hansen, H. S. (2023). Analytics of historical human migration patterns: Use cases of Amsterdam and Copenhagen.
Stonawski, M., Georgati, M., Crisci, M., Wissen, L., & Brzozowski J. (2023). Report on migration and residential segregation in European cities – now and future – analysis and projections of immigration and spatial distribution of immigrants. D6.4
Georgati, M., Stonawski, M. Keßler, C., & Hansen, H. S., A comparative study of spatio-temporal segregation patterns with gridded population data in Denmark, the Netherlands, and Poland.

Fig. 1.2: Structure of the PhD dissertation (left), research design (middle) and contributions to papers and reports (right). Dark blue highlights primary contributions (first author), medium blue secondary contributions (co-author), light blue contributions to a project deliverable and white contributions in progress.

2

Background and state of the art

Background and state of the art

This section outlines the early research that gradually guided the selection of methodology to be followed in the study. It starts by introducing important concepts related to population projections in section 2.1, leading to advances in migration modelling in section 2.2 and population projections at regional and gridded levels in sections 2.3 and 2.4. Each subsection briefly describes the approach adopted by FUME in the corresponding steps. Section 2.5 introduces various types of ML models relevant to the project's context and discusses the role of interpretability in ML development.

2.1 Population projections

Population projections describe the future size and composition of the human population [88]. The projections may refer either to the total population or to its breakdowns by demographic characteristics (e.g. age, gender, education), or they may distinguish between individual components of growth cohorts [134]. Fertility, mortality and migration [in-migration and out-migration; 30, 14, 134, 155] are the major components of population change and define the final size of the population of a geographical area at a given future date [106].

In order to produce population projections, demographers make assumptions about the levels of fertility, mortality, immigration and emigration at different time steps over the projection period [88]. Accounting for the uncertainty and likelihood associated with these assumptions is essential to ensure the appropriate use of the projections. The two major approaches used to characterise the uncertainty in these assumptions are scenarios and probabilistic projections [107]. A more detailed discussion of the former follows, as it is the most commonly used class and the one used in FUME.

Accounting for uncertainty with scenarios

The scenario-based approach relies on narratives to make assumptions in order to cover a range of possible futures and to reduce the degree of uncertainty [107]. Population projection methodology adopted the use of scenarios in the early 1990s to cover alternative calculations and to describe possible sequences of events, some of which are unlikely to occur [84]. Each scenario assumption is usually quantified through various components, such as economic growth or technological development. It may also refer to different spatial scales, from continents to countries and cities. The quality of the scenario is critical to the final product, as the numerical outcome is highly dependent on the underlying narratives.

A range of scenarios for population, socio-economic and climate development have been published, with the Special Report on Emissions Scenarios [SRES;

102], the Representative Concentration Pathways [RCPs; 100] and the Shared Socioeconomic Pathways [SSPs; 108] providing widely used narratives with qualitative differences. The SSPs create a new scenario framework for climate change that presents possible alternative future trends for the world and large world regions, expressed in five narrative storylines and their corresponding sets of quantified development measures [108]. The SSP2 '*Middle of the Road*' was the benchmark against which the FUME scenarios were developed.

Methods for population projections

Apart from accounting for the uncertainty in the assumptions, the selection of a suitable methodology plays a significant role in the development of reliable population projections. O'Neill *et al.* [107], Booth [14] and Smith *et al.* [134] provide comprehensive reviews of various methods used for population projections. They emphasise that the choice of method is influenced by numerous factors, with data availability, the objective of the projection and the researcher's expertise and judgement being the most critical. However, it is the diversity of users' needs that specifies choices around the spatial scale, time frame of projections and variables that consequently drive the selection of the suitable method of projection.

Smith *et al.* [134] identify four primary methods for population projections: the cohort-component method, trend extrapolation, structural models and microsimulation models. Of these, the cohort component method is the most established for long-term global projections [107], applied by academic demography, national governments and international organisations such as the United Nations (UN) and the World Bank [18].

This method accounts for the components of population growth – births, deaths and migration – separately by dividing the population into groups by age and sex and making projections for each group independently. One of the key aspects of this method is its strong dependency on the initial size and age structure of the population, as well as age-specific fertility, mortality and migration rates to forecast the future size and age structure of the population at various time points [107]. The UN and Eurostat are the main sources of world or European population projections by age and sex [39, 148].

The multistate methodology is an extension of the cohort component framework in which additional characteristics of the population, such as educational attainment or household composition, are taken into account to measure and project changes in status over the life course of individuals [114]. Original work includes multistate projections accounting for the place of residence [123] or place of birth [114] conceiving of states primarily as geographical units. However, states generally refer to subgroups of the total population [86]. An interesting instance of the multistate methodology is presented by Lutz and Goujon [86] who produced a '*complete matrix of the composition of the population by age,* *sex and educational attainment for different points in time*' [86, p.324].

The multistate method has been extended to the '*multi-dimensional mathematical demography methodology*' [72] providing national population projections based on the five SSP assumptions on fertility, mortality, migration, age and education. IIASA/Wittgenstein Centre (WIC) used a multidimensional demographic cohort-component population projection model to estimate population stocks by sex, age and education at the national level [85, 87]. An adapted version of this model including CoB was used in FUME.

Population projections in FUME

In FUME, six global migration scenario narratives were developed relying on an expert survey – recruiting academic researchers –, a Delphi survey – recruiting policy makers and advisors to policy makers –, while also taking into account findings on push factors in origin countries and the impact of the COVID-19 pandemic [159]. These scenarios were based on the SSPs with the SSP2 '*Middle of the Road*' as the starting point. The scenario assumptions were quantified through variations in national GDP projections at global scale with the estimates from Koch and Leimbach [76] used as benchmark.

The primary objective of FUME was to gain insight into international migration patterns and flows under different migration scenarios in Europe [71]. To accomplish this, KC *et al.* [71] developed a multidimensional demographic model to disaggregate the population by age, sex, education and CoB [71]. Recognising the influence of educational attainment on migration rates in addition to age and sex, KC *et al.* [71] used the model previously developed by the IIASA and incorporated some of its assumptions regarding fertility, mortality and educational transitions. KC *et al.* [71] extended the IIASA model for FUME by introducing CoB as an additional dimension.

An important element in population projections is migration modelling. Due to its significant role and its significantly less predictable nature compared to fertility and mortality, migration modelling is discussed in the next section, which concludes with a description of how migration flows were modelled in FUME and then integrated into the population model. The migration component was developed by the PIK and replaced the previous assumptions on linear extrapolation of historical migration rates within the IIASA population projection model.

2.2 Migration modelling

The estimation of migration flows, driven by the assumptions made in migration modelling, plays a crucial role in population projections. This subsection provides an overview of the evolution of migration modelling throughout history and concludes by highlighting contemporary approaches to projecting migration flows, including the FUME model.

Migration modelling has long attracted the attention of geographers under the umbrella of spatial interaction and distance decay, attempting to describe spatial flows mathematically [46, 105]. Exploring the literature around the distance effect, H.C. Carey [19] observed in his Principles of Social Science that the amount of interaction between two cities is proportional to their population size and inversely proportional to the distance between them [11]. In addition to distance constraints, Ravenstein [117] introduced socio-economic factors as fundamental issues for population movements and formulated his laws of migration in one of the earliest scientific studies of internal migration. Ravenstein [117] analysed '*lifetime place-to-place*' migration and net migration and underlined the *'spatial variations in economic opportunities*' through his comparison between male and female migration and their choice of residence location respectively [57, p.6-7]. Overall, his approach formed the basis of the *gravity model* of spatial interaction and introduced crucial issues that continue to affect modern societies.

In 1940, Stouffer [141] introduced the laws of intervening opportunities, which break the link between mobility and distance and establish a new relationship between opportunity and mobility. Specifically, it is proposed that '*the number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities*' [141, p.846]. This concept is very much in line with modern societies, where employment and education are two of the most popular reasons for migration, while technological developments have made transport cheaper and faster.

Approximately the same period, Zipf formulated the *Gravity Law*, which states that *'the number of persons moving between two communities with populations P1 and P2 respectively and which are separated by the shortest transportation distance D will be proportionate to the ratio* $P1 \times P2/D$' [160, p.677]. Distance is the common theme between Ravenstein, Stouffer and Zipf, '*though its functional effect on movement*' differs in the latter two models [11, p.5].

Similarly, Rich, discussing the gravity model, states that '*the volume of interaction between two cities is a positive function of their population sizes and an inverse function of the distance between them*' and presents the '*frictional effect*' of movement between two areas [119, p.4]. This '*frictional effect*' or, differently put, the '*difficulty*' of movement is proportional to the geographical distance between them, and inversely proportional to the likely frequency of interaction between the populations of the two areas, and results from various factors, such as transport costs.

Evidently, space is a key element in migration studies, and spatial structures significantly influence migration patterns [27]. However, Cushing and Poot [27] argue that the vast majority of empirical research fails to include any aspect of space and that '*omitting spatial structure may seriously bias most empirical conclusions*' [27, p.325]. According to them, distance is the key to studying demographic or place characteristics, types of migration and temporal changes in migration flows.

Distance, however, cannot be the sole determinant of migration flows, especially in today's societies where the drivers of migration have become more complex and travel has become more affordable, rapid and convenient. Today, the attractiveness of origin and destination, including socio-economic factors, employment opportunities, infrastructure improvements, crime or even climate change, as well as personal preferences, are more important measures to take into account when studying migration flows than distance.

In this direction, various modelling approaches have explored different factors to explain migration patterns and make projections for the future. These factors include past migration rates [125, 16], global and national wealth [2], changes in education [36], diaspora or return migration [120]. The significance of these factors has been adjusted according to different scenarios.

The existing literature on utilising ML for predicting human migration flows is limited though. Robinson and Dilkina [121] conducted a comparative study where they introduced two newly developed ML-based models alongside two traditional models (i.e. a gravity and a radiation model; 121). These ML-based models were employed to predict migration flows between counties within the United States (US) and between countries worldwide. The study utilised two sets of features and demonstrated that the ML-based models outperformed the traditional models across various evaluation metrics. The ML models showed greater flexibility by incorporating a wider range of input features in a non-linear way and they could be easily customised for different spatial levels.

Data collection is an essential component of migration modelling. However, despite increasing availability, data on international migration flows remain incomplete, inconsistent across sources and insufficient to provide a comprehensive and holistic view of actual global migration flows. To address this, several studies have proposed methods to overcome data inconsistencies and incompleteness, including approaches suggested by Abel and Cohen [3], Azose and Raftery [9], Abel and Sander [4] and Abel [1].

Migration modelling in FUME

FUME introduced a dynamic model of global bilateral migration that incorporated important mechanisms such as the impact of diasporas, including transit migration flows, poverty constraints based on observed emigration rates in poor countries, and return migration based on migrant stocks. The model was calibrated using a global dataset of bilateral flows, demonstrating its ability to accurately capture past migration levels and trends while projecting future migration flows [67, 120].

In order to capture the complexities arising from education and age, the model was expanded to include multiple dimensions of population heterogeneity [71]. As a result, the final population projection model employed in FUME incorporates a more mechanistic representation of migration compared to the SSP projections. This enhanced model responds dynamically to various scenarios, such as changes in national income levels or educational attainment, allowing for a more nuanced understanding of migration patterns [71].

2.3 Regional projections

Whereas national and global projections can cover up to a century, limiting the output variables to age and sex [107], local projections are typically limited to short-term projections (up to 10 years), including a wide range of characteristics such as education, marital status or urban residence. This subsection discusses instances of sub-national and regional projections.

As with global and national population projections, Eurostat has published projections of the future size and structure of the population at regional level in Europe (i.e. EUROPOP2008; 55, EUROPOP2019; 40). Such projections can be used to examine different policy frameworks, demographic trends and migration patterns in European regions and the implications for regional competitiveness and European cohesion [118].

IIASA has been implementing the multistate methodology for sub-national projections. Questioning the impact of socio-economic heterogeneity and urbanisation patterns on population projections, KC *et al.* [73] demonstrated sub-national population projections for India. Specifically, they developed a five-dimensional model for the population of India, disaggregated by state, rural/urban place of residence, age, sex and education. They showed that education has a significant impact on fertility and mortality, suggesting that if the pace of education expansion observed at the time continues, '*India will rapidly catch up with other more developed countries in Asia*' [74, p.8328]. An earlier approach by IIASA presented regional population projections for China, motivated by China's regional diversity and its increasing integration into the world economy [146].

The MULTIPOLES model [78] incorporated a cohort-component, multi-regional and supra-national approach, allowing simultaneous projections for all countries. It also accounted inter-regional internal migration, inter-regional migration within Europe and international migration to and from regions outside Europe [118]. Similarly, the cohort component technique was used to produce small area population projections at the census tract level in Clark County, Nevada [143] and at the municipal subdivision level in Denmark [59]. The Danish model incorporated changes in dwelling structures and divided immigrants into groups by CoO to provide numerical forecasts of future migration and refugee flows [59]. Another approach used regression trees and downscaled five age groups using US counties based on historical differences [142].

Regional projections in FUME

In FUME, the sub-national projection model bridged the gap between the global and local models. This model incorporated several important features, such as a bi-regional level, comparing the city region with the rest of the country, and a multistate model with various dimensions, including age, sex, region, CoB and educational attainment [153]. Furthermore, the data used in these models account for the dynamics of inter-regional transitions, such as migration from the rest of the country to the urban region, as well as educational changes [153].

For the sub-national projections of Amsterdam and Copenhagen, the model utilised a comprehensive five-dimensional cross-classification of the population [153]. However, the models for Kraków and Rome were constructed based on more limited information. The available data included national totals and city-specific information. By subtracting the information for the rest of the country, the model was able to generate projections for these cities [153]. Proceeding to the local modelling, the subsequent subsections focus on research instances on gridded population projections, the associated challenges and the added value that advancements in ML bring to this field.

2.4 Gridded population projections

The previous subsection presented examples of population projections for subnational and small area regions, extending the explored dimensions beyond age and sex. Nonetheless, cohort-based and multi-dimensional population projections become particularly challenging when applied to these smaller spatial scales. This is primarily due to the need for sufficient data on dimension-specific (e.g. age, migration) demographic rates for each individual spatial unit, covering both historical patterns and future projections [142]. Such data on current conditions are typically lacking or difficult to obtain at smaller spatial scales. Moreover, making future projections for local areas involves numerous assumptions, which increase the challenges with factors that exhibit spatial and temporal variations, such as migration [142].

Despite these difficulties, there is often a need for *finer spatial resolution*, especially when integrating social and environmental processes. As a consequence, many studies have reduced or simplified the dimensions of their analyses in order to increase the resolution, producing spatially explicit population projections at resolutions between 1 kilometre (km) and 0.5°. They focus on total populations and densities or aggregated groups, such as urban or rural populations, rather than explicit population sub-groups (e.g. age or migrant groups). In global population projections, Jones and O'Neill [69] provided projections for total, urban and rural populations at the national level for all countries in the world. Similarly, in the context of the US, Jones and O'Neill [70] focused on spatially explicit population scenarios for internal migration, utilising a potential model that estimates the attractiveness of different locations. The model downscaled aggregated population projections to grid cells and incorporated socio-economic assumptions. The model of Zoraghein and O'Neill [161] captured greater spatial variations and provided a basis for integrated environmental analysis with demographic uncertainty. All these studies expanded the gravity-based modelling approach and incorporated the SSP narratives to explore alternative development patterns.

Instead of using variants of the gravity model, alternative approaches have suggested geosimulations to project future population distributions at finer resolutions. Gao [49] downscaled the 1/8-degree projections produced by Jones and O'Neill [69] to 1 km grid cells with a simple approach employing the 2000 population count map from the Global Rural-Urban Mapping Project [GRUMP; 10]. Motivated by extreme climate change as a potential reason for massive migration waves, Keßler and Marcotullio [75] predicted the population at 1 km resolution, distinguishing between urban, suburban and rural cells. Using a weighted surface, McKee *et al.* [94] projected population with the LandScan [35, 104] dataset – a high-resolution population distribution – as a baseline.

Overall, while gravity-based approaches allow projection of the population, they can be time-consuming and require detailed datasets for many demographic dimensions, which are typically not readily or openly available at small area regions or gridded scales. Simplifying the analysis to aggregated groups, such as urban/rural populations, allows researchers to work with available data and address broader trends or patterns.

Benefits of machine learning for fine-grained population estimates

Geosimulations and recent advances in ML and GeoAI offer advantages in capturing the complexity of population dynamics, enabling estimates at finer spatial and temporal resolutions. Access to an ever-growing amount of topographic and socio-demographic data coupled with these new methods is also reshaping how population projections and estimations are being developed. The emergence of methods that combine spatial analytics, ML and simulation paves the way for novel solutions to spatial interaction problems using rich datasets [45]. Such approaches can be used for direct future projections, for creating new datasets or for completing existing ones and increasing their accuracy and spatio-temporal granularity.

Examples of direct future projections for gridded population distributions might not be available yet in the literature, but ML has been applied for projecting built-up land based on time series of remote sensing observations [50] and for estimating existing urban structures [7] and population density [44, 122, 115, 132]. These approaches have demonstrated that ML can assist the extraction of information from unstructured, remote sensing data and provide effective solutions for socio-demographic issues.

Other studies have employed ML models in population density disaggregation approaches such as the Global Human Settlement Layer [GHSL; 128], the Gridded Population of the World [GPW; 21, 22], WorldPop [158] and Monteiro *et al.* [99, 97, 98]. In such cases, the estimation of historical distributions is possible, and with the rapid technological advances, the progress in both temporal and spatial resolution is remarkable. A few years ago, the publicly available GHS-POP datasets were limited to a resolution of 250 m with controversial accuracy [77]. They are now available at 100 m resolution with significantly improved accuracy. These approaches provide more robust and accurate fine-scale gridded population maps, while reducing the modelling complexity, and contribute to the deeper understanding of processes related to climate change, facilitate the mapping of socio-environmental variables and resource allocation, and support the Sustainable Development Goals at the global level [130, 126].

Different ML models offer unique strengths and limitations for population projections or estimations. Choosing the right model depends on the nature of the problem, data characteristics and desired outcomes. To make an informed decision, it is important to understand the differences, advantages and limitations of various models. The following section explores different types of ML models, their benefits and presents applications relevant to the research context.

2.5 Machine learning models

Neural spatial interaction models were already proposed as an alternative to the fully constrained gravity models to capture patterns in data in the 1990s [13]. The gravity artificial neural network (GANN) approach proposed by Black [13] is one of the early efforts in implementing neural network models in demographic research and revealed its great potentials offering higher flexibility and accuracy with less coefficients. This field of research, later known as *GeoAI*, has now greatly progressed because of the large-scale availability of high-quality data and the advances in the technologies processing these data. These advancements have effectively shown that models that consider spatial details perform significantly better when applied to spatial data than generic ones [68].

This subsection presents different types of ML models suitable for regression problems, where the goal is to predict a continuous output variable (e.g. population density). The exemplary applications are selected taking into consideration the context of the project and the instances discussed earlier. The subsection concludes with the evaluation of model performance and the role of interpretability in ML development.

The examples consist of both supervised and unsupervised learning, although unsupervised learning, by its nature, is not typically applied directly to regression problems. Supervised learning involves training a model using labelled data, where input data are paired with corresponding output labels or target values [58]. The model learns from these labelled data to make predictions, such as estimating the population count in each grid cell. Unsupervised learning, on the other hand, utilises unlabelled training data, allowing the system to learn without a teacher [58] such as for clustering (i.e. for typologies of residential built-up areas; 91).

Some algorithms commonly used for regression tasks are the following:

1. **Linear regression**: Linear regression establishes a linear relationship between input features and a continuous target variable [61]. It can be used for population projections or estimates by establishing a relationship between population size and relevant predictor variables, such as historical population data, urbanisation rates or built-up areas. As the predictions is based on the linear relationship of the population size and the predictor variable, the prediction accuracy depends on the nature of the training data and the presence of any non-linear relationships.

2. **Decision trees**: Decision trees are versatile ML algorithms that recursively split the data based on features to make predictions [58]. They are suitable for handling a wide range of predictor variables in population estimations and can capture non-linear relationships and interactions between variables, which is beneficial when dealing with complex population

dynamics. Additionally, they offer interpretability, allowing analysts to understand the reasoning behind the population estimates by examining the decision paths and splitting criteria of the tree. However, they suffer from overfitting if the model becomes too complex or the data are noisy.

3. **Random forest (RF)**: RFs [15] are ensemble models consisting of multiple decision trees. They combine by averaging the predictions of individual trees to make more accurate predictions [61]. They are advantageous for population estimates as they can capture non-linear relationships and handle interactions between variables, mitigating overfitting and providing robustness against noisy data. They also allow for feature importance analysis, enabling identification of the most influential predictors for population estimates.

4. **Gradient boosting (GB)**: GB is another ensemble method that combines multiple weak predictive models to create a strong predictive model [58]. The model is trained sequentially, learning from the errors of the previous model and improves over iterations. GB can effectively handle complex relationships in the data and achieve high predictive accuracy. It also allows for the interpretation of feature importance, as it quantifies the contribution of each predictor variable to the overall model. Lastly, it handles missing data and outliers by considering their impact during the training process.

5. **Artificial neural networks (ANNs)**: ANNs are powerful models inspired by the structure and function of the human brain, consisting of interconnected nodes (neurons) organised in layers [58]. Deep learning (DL) is a subset of neural networks that focuses on learning hierarchical representations of data. DL models consist of multiple layers of interconnected nodes, enabling them to learn complex patterns and features automatically [82]. Specialised neural networks are the following:

   (a) *Convolutional neural networks (CNN)*: CNNs are designed to process grid-like data [56], such as images and rasters. They use convolutional layers to extract relevant features and are widely used in computer vision [79]. CNN models have been used for population estimations using Earth observation data [122, 115, 44] and in disaggregation approaches [98].

   (b) *Generative adversarial networks (GAN)*: GANs consist of two interconnected, competing neural networks – the generator and the discriminator. They are used to generate new, synthetic data that resembles a given training dataset. They are able to generate realistic and high-quality samples in various domains, such as images [82]. Albert *et al.* [7] generated hyper-realistic urban patterns through an unconstrained GAN with aim the apprehension of urban diversity forms around the globe.

(c) *Recurrent neural networks (RNN)*: RNNs are designed for processing sequential data [82], such as time series. They have feedback connections that enable them to retain information from previous steps, making them suitable for processing sequences and '*predicting the future*' [58, p.381]. The *long short-term memory* [LSTM; 62] is a type of RNN architecture designed to capture long-term dependencies and handle sequential data. It addresses the limitations of traditional RNNs, which struggle with retaining and utilising information over longer sequences [58]. It incorporates specialised memory cells and gating mechanisms to selectively remember or forget information at different time steps [58]. It consists of the input gate, the forget gate and the output gate [58]. The availability and quality of historical data play a crucial role in the accuracy of population estimates. Gathering and preprocessing reliable and comprehensive historical data is essential for the success of RNN-based population estimation models. They require careful hyperparameter tuning and regularisation techniques to prevent overfitting and improve generalisation. Gao and O'Neill [50] developed the Spatially-Explicit, Long-term, Empirical City developmenT (SELECT) model for projecting global built-up land based on time series of remote sensing observations.

The evaluation of a ML model is an essential step following its development to assess its prediction accuracy and overall performance. This can be achieved by using common metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), R-squared (R^2) or mean absolute percentage error (MAPE) in supervised learning, and visual and comparative analysis and cross-validation in unsupervised learning [43]. These metrics provide quantitative measures of model performance and prediction quality, while a combination of multiple metrics can provide a more comprehensive understanding of the model's capabilities. In geospatial modelling, it is also important to include spatial measures to examine the error distribution, as explored in Paper B (sec. 9). RF, GB and CNN are the primary models employed in this PhD project, as shown in Papers A, B and C (sec. 8, 9, 10).

The role of interpretation in machine learning models

The interpretation of the reasons behind the model's predictions has not received the same attention as the development of increasingly more accurate models. While performance is crucial, it is important to recognise that single metrics alone cannot explicitly and fully describe the modelling process. Specifically, Molnar [96] wonders:

'*If a machine learning model performs well,*
why do we not just trust the model and ignore why it made a certain decision?'

Driven by the need to understand and explain the factors that contribute to the model's outcomes, the remainder of this subsection focuses on the significance of interpretation. By delving into interpretation, this study aims to shed light on the holistic understanding of the modelling process beyond mere performance metrics.

The concepts of interpretability and explainability have recently become very important and of increasing interest to the research community [20]. In terms of terminology, some generally refer to '*interpretable machine learning*' as '*methods and models that make the behaviour and predictions of machine learning systems understandable to humans*' [96]. Others distinguish the two terms with explainability encompassing the techniques that assist users understand the reasons for a model's behaviour, and interpretability referring to the user's ability to understand and analyse the model's prediction and what the model decided while making the prediction [31].

There are many reasons to interpret and understand why a prediction was made and what factors influenced that prediction. First and foremost, interpretability promotes transparency, allowing us to understand the reasoning behind the model's predictions or decisions [20]. It promotes understanding of the underlying factors and variables that contribute to the model's output. It builds social acceptance, trust and confidence in the model [96], as stakeholders can verify that its predictions are based on reasonable and fair judgements. It also ensures safety and accountability, especially in cases where the model's outputs have significant impacts on individuals or society [96]. Lastly, human curiosity and the desire to explain the behaviour of the model and to make sense of the world or to detect bias drive the demand for increasing the interpretability of ML models [96].

Moreover, interpretability provides valuable insights into feature importance. Understanding which variables or features have the most significant impact on the model's predictions promotes a deeper understanding of the underlying phenomena, enabling further investigations or informed actions. It is noteworthy that especially scientific fields such as migration, that were previously primarily oriented towards qualitative research, have witnessed a shift towards quantitative methods. ML and large datasets have proven instrumental in solving numerous challenges and generating new knowledge [96]. In this context, interpretability helps to extract and build on the existing knowledge captured by the model, facilitating data-driven research and the ability to explain phenomena through empirical analysis.

ML algorithms can be interpretable, also called white-box models, or non-interpretable, also called black-box models [31]. White-box models provide a means to understand the prediction mechanisms and insights into how and why the model makes certain decisions (e.g. linear regression models). Black-box models do not reveal their internal mechanisms and their decisions cannot be explained by their parameters (e.g. neural networks) [96]. Black-box models generally have better predictive performance than white-box models [31], but the trade-off between predictive performance and uncertainty about how the system works depends on the nature of the problem.

3

Case study areas

Case study areas: data and methods

A large part of this research involved the collection, preparation and visualisation of data for the selected case studies. Various geospatial data (e.g. points of interest, infrastructure locations) were collected by scraping online sources such as web maps and open data portals, in addition to processing data provided by local authorities. Depending on the availability and quality of the data, it was sometimes necessary to combine data from different sources to create a more comprehensive and holistic dataset. For instance, datasets on existing and future urban development plans were created or supplemented utilising static maps and descriptions from local plans to be used in training for future projections.

Data preparation included pre-processing to clean, normalise and transform the data into different formats and types depending on the end use and purpose. Visualisation was also important to analyse and present data and results. Various types of maps and graphs were produced during this process, including choropleth, bivariate or interactive maps, and scatter, bar or stacked plots.

The FUME proposal suggested four case studies to examine migration at the local level in destination cities in Europe. The proposal selected representatives from different regions of Europe: Copenhagen (North), Amsterdam (West), Rome (South) and Kraków (East). Figure 3.1 illustrates the location of the selected case studies together with the share of international migrants at country level in the EU. Darker shades of green indicate a high concentration of migrants relative to the total population, as in the cases of Switzerland and Luxembourg, where more than 20% of the population is foreign born. On the EU's eastern borders, in light blue, migration is limited to less than 5%, as in the cases of Bulgaria, Romania and Poland.

The initial intention of the project was to train a ML model on the historical distribution of migrant populations by CoO at a resolution of 100 m grid cells. However, as presented in section 1.2, the uncertainty of collecting the required amount and quality of data posed challenges to this approach. Despite the difficulties, the FUME partners managed to collect a large amount of data on the distributions of migrants at the targeted resolution for almost all the selected case studies. However, there were differences in definitions, availability, accessibility and quality among them. Table 3.1 summarises the population datasets collected by the FUME team, their sources, coverage, time frame, type of access, resolution and definition of migration background.

The maps in Figure 3.2 show the distribution of migrants in relation to the total population based on the collected data. The visualisation employs 1 km grid cells within the municipal boundaries for the examined areas in 2020. Notably, the Danish map reflects the distribution of migrants in 2018, as 2020 data for Greater Copenhagen were not available when these maps were created. Moreover, the $\mathrm{CBS_g}$ dataset substitutes the $\mathrm{OIS_g}$ dataset due to limited

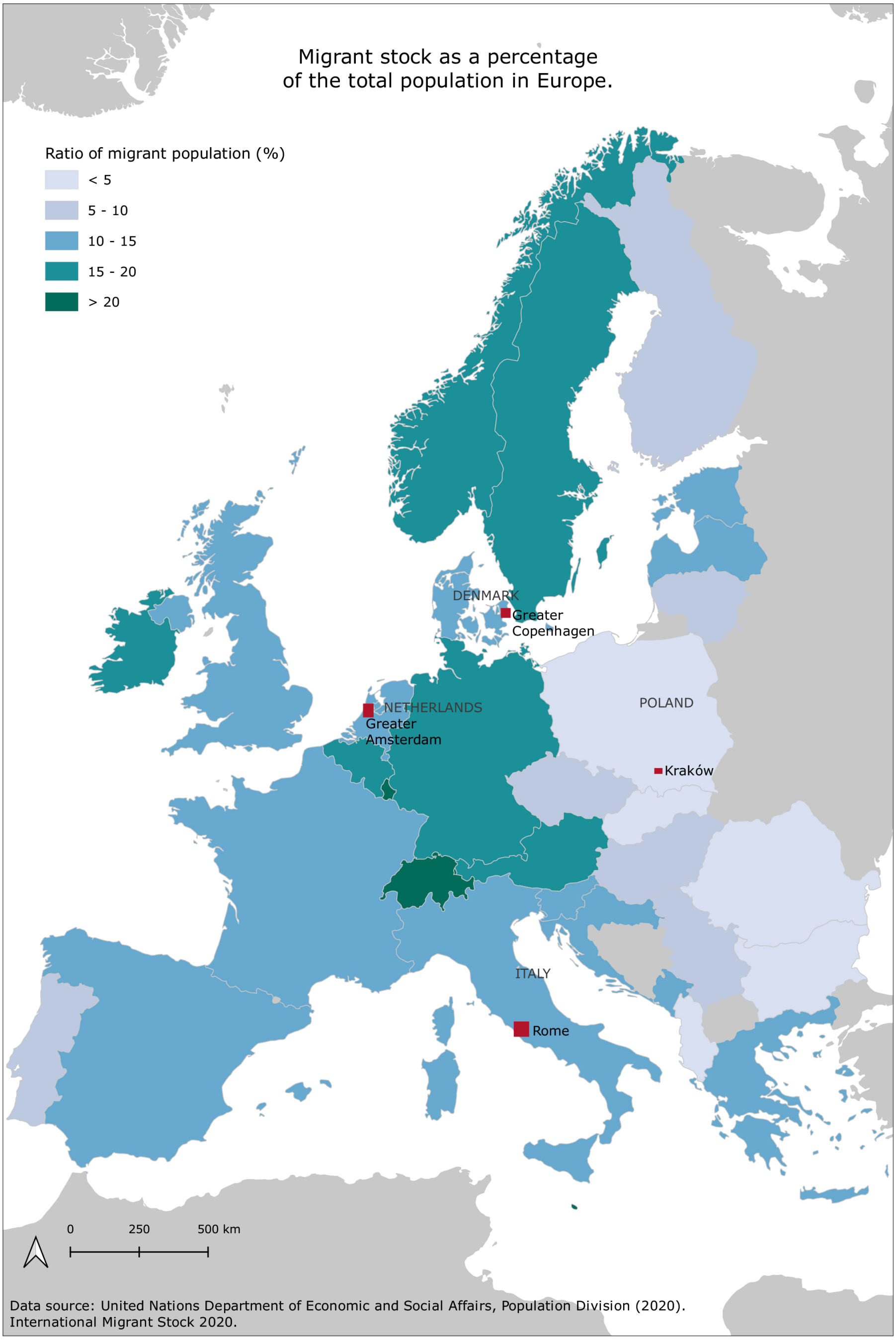
Table 3.1: Summary of the available population datasets by source, spatial and temporal coverage, type of access, spatial resolution and migration background breakdown. Abbreviations: RA – restricted access, NDRD – non-redistributable data, MB – migration background (Western – non-Western), CoO – country of origin, CoC – country of citizenship, RoC – region of citizenship, E – only even years, A – all years.

Dataset	Source/Coverage	Time frame	Type	Resolution	Class
$\mathrm{OIS_g}$	Mun. of Amsterdam/ Mun. of Amsterdam	1992–2020 (E)	NDRD	100 × 100 m	CoO
$\mathrm{OIS_{g1}}$	Mun. of Amsterdam/ Mun. of Amsterdam	1992–2020 (E)	NDRD	1 × 1 km	RoO
$\mathrm{CBS_g}$	CBS/Netherlands	2015–2022 (A)	Open	100 × 100 m	MB
$\mathrm{DST_g}$	DST/Greater Copenhagen	1990–2020 (E)	RA	100 × 100 m	CoO
$\mathrm{DST_{g1}}$	DST/Greater Copenhagen	1990–2020 (E)	NDRD	1 × 1 km	RoO
CASPAR	Mun. of Kraków, Voivodship Office, FUME team/Mun. of Kraków	2013–2021 (A)	NDRD	100 × 100 m	CoC
CNR	Mun. of Rome, ISTAT, FUME team/Mun. of Rome	2001, 2011, 2015–2020 (A)	NDRD	Census tracts	RoC
$\mathrm{CNR_g}$	Mun. of Rome, ISTAT, FUME team/Mun. of Rome	2020	NDRD	100 × 100 m	CoC

spatial coverage in the Municipality of Amsterdam. The aggregation methods vary, including either summing the total and migrant populations at 1 km grid cells and estimating the percentage when available (Greater Copenhagen, Kraków, Rome) or averaging the share of migrants (Greater Amsterdam). All data processing, modelling and visualisations were performed in the ETRS89 Lambert Azimuthal Equal-Area geospatial coordinate system (EPSG: 3035).

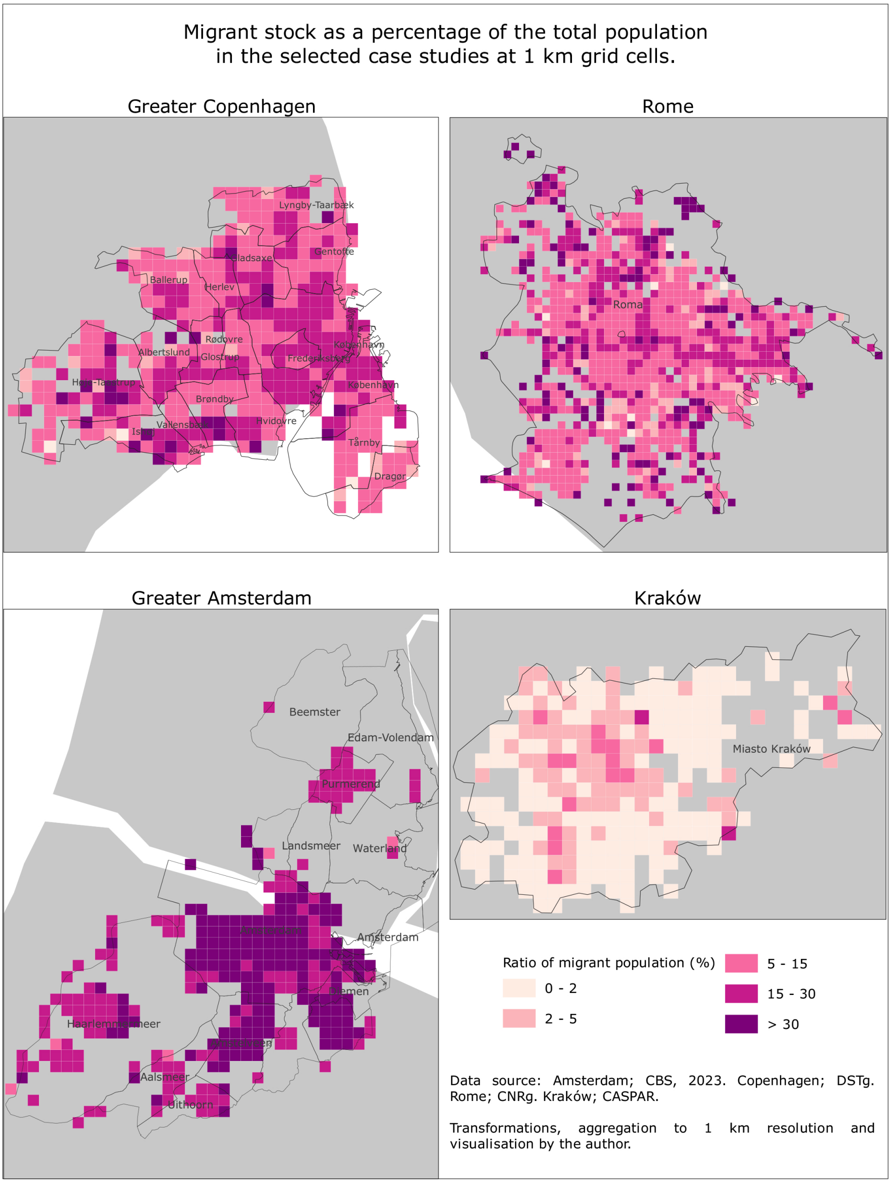
The maps display the different concentrations and distributions of migrants among and within the case studies. In Greater Amsterdam, most cells are in dark purple with a concentration of migrants above 30% of the population, while Greater Copenhagen and Rome show a more varying distribution with most cells in light purple and pink. These colours indicate concentrations between 5 and 30%. In contrast, in Kraków the concentration is low, with a few exceptions where the concentration exceeds 5%.

These diversifying patterns within the case study areas are briefly described below in the following subsections. Specifically, sections 3.1 - 3.4 provide descriptions of each case study area, discussing their historical significance as migration hubs. They also present aspects of the population and ancillary data collected for the study. These sections also address the challenges associated with the corresponding datasets, focusing primarily on the advantages of each city as a case study, as well as the limitations arising from data quality and availability, processing environments and variations in definitions.



Data source: United Nations Department of Economic and Social Affairs, Population Division (2020). International Migrant Stock 2020.

Fig. 3.1: Ratio of foreign-born population to total population in European countries in 2020. The map highlights the countries with the highest concentration of migrants in dark green and the location of the case studies in Europe in red.



Migrant stock as a percentage of the total population in the selected case studies at 1 km grid cells.

Greater Copenhagen

Rome

Greater Amsterdam

Kraków

Data source: Amsterdam; CBS, 2023. Copenhagen; DSTg. Rome; CNRg. Kraków; CASPAR.

Transformations, aggregation to 1 km resolution and visualisation by the author.

Fig. 3.2: Ratio of foreigners to total population in the four selected case studies at a resolution of 1 km. The maps show the higher concentrations of migrants in purple, in the western part of Greater Copenhagen, in the urban periphery of Rome and in the central part of Greater Amsterdam and Kraków. The maps show the boundaries of the corresponding municipalities.

3.1 Amsterdam

Due to its colonial past, Amsterdam has always been an important migration hub, attracting migrants from the former colonies. Meanwhile, its current economic and technological advancement continues to attract a diverse workforce from Europe and around the world. As a result, Amsterdam is an interesting case study to explore long-lasting migration patterns and their potential future prospects. In 2020, 39% of the city's population had a foreign origin, with the largest groups coming from other high-income countries, Turkey and Morocco, the former colonies and the countries of Middle East and North Africa (MENA), as estimated from $\mathrm{OIS_g}$ dataset.

A large influx of guest workers from Morocco and Turkey was attracted between 1960 and 1970, followed by a subsequent wave of family reunification [138]. In 1975 Suriname, a former colony, became independent and many Surinamese migrated to Amsterdam and settled in the available housing in the newly built high-rise district of Bijlmermeer [138], in Zuidoost, where the presence of migrants remains high, as shown in Figure 3.3. Although the influx stopped in the 1980s, the Surinamese are still one of the largest migrant groups in Amsterdam. However, their population has been declining since 2000, partly due to a '*large exodus*' from the city to the countryside, since they formed a well-educated middle class, and partly due to return migration [38, 154].

Since the 1990s, migration has been characterised by labour migration from the EU and other high-income countries, or by refugees from the former Yugoslavia, the Middle East and African countries. A sharp increase in migrants from Central and Eastern European countries, especially Poland, followed their accession to the EU. After 2010, a large number of highly skilled migrants from the Western EU started to settle in Amsterdam [38]. Amsterdam is also an attractive destination for international students [138].

As for its urban development, after a long phase of urbanisation until the 1960s, the residents of Amsterdam moved out of the city and into newly built suburbs. This led to a deterioration of the urban environment [138]. The 3rd and later the 4th Memorandum of Spatial Planning in 1974 and 1988 [113] shifted the focus from the suburbs to the cities and aimed to densify the urban environment where services, communication, urbanisation and globalisation would play a key role in the development of the city until today [138].

An additional advantage of this case study is the availability of extensive statistical data on demographics, housing, infrastructure and proximity to facilities, which are openly available either at various administrative levels or at grid cell level [23, 51, 103]. In these datasets, the migrant groups are aggregated either by Western or non-Western migrant background or by the more recently established classification by continent [137]. To comply with confidentiality restrictions, the openly available gridded datasets have been processed by imputing the missing values for cells with a low number of inhabitants or households, thereby reducing the accuracy of the dataset.

Since the initial goal of the project was to train a ML model on the historical distributions of selected groups, replacing low value cells with missing values was considered unsuitable. Such replacement adversely affects the interpretation and accuracy of the data, introducing uncertainty and bias. Additionally, the aforementioned classifications were not appropriate for the specific study due to its focus on more specific groups at either country or region level. Due to these reasons, we accessed confidential population and building data from the Municipality of Amsterdam that contained information on the distribution of the population by CoO/CoB without replacing the lower values. Although the accessed data were not used to train a time series model, they served as the ground truth dataset for evaluating the disaggregated distributions at the pixel level.

For Papers A and C (sec. 8, 10), the study focused on the Municipality of Amsterdam as a case study, utilising the gridded dataset with a resolution of 100 m. In Paper D (sec. 11), the study aggregated the aforementioned gridded dataset to a resolution of 1 km, which was used for the analysis (OIS_{g1}).

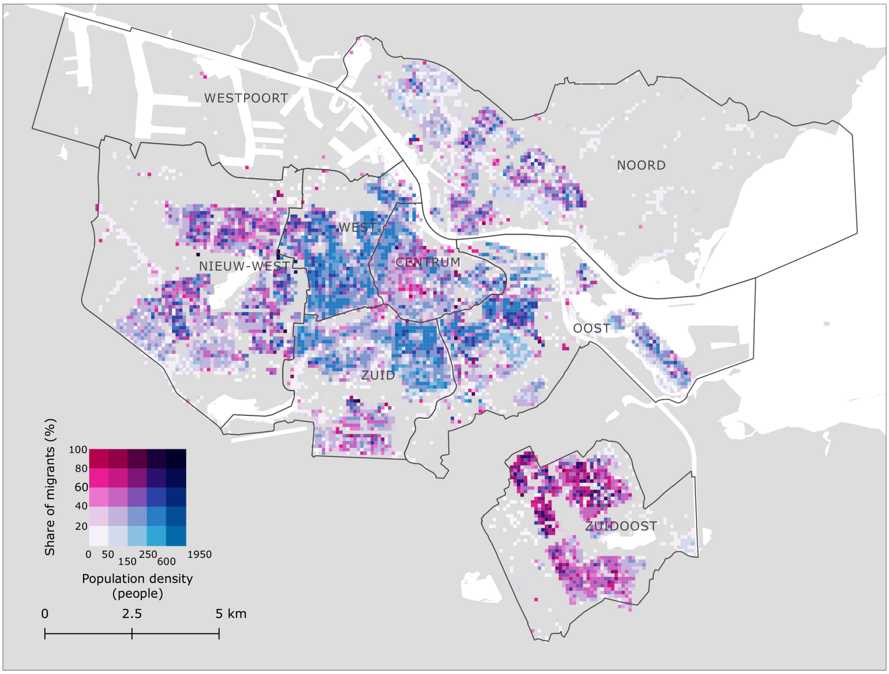


Fig. 3.3: Population density and share of migrants in the Municipality of Amsterdam in 2018. Areas highlighted on the map in pink for their high migrant concentration: Zuidoost for Suriname and non-Western migrants, Nieuw-West for those from the MENA, and the Centrum for EU migrants. Data source: OIS_{g}.

3.2 Copenhagen

The population composition in the metropolitan region of Copenhagen has significantly changed over the last 30 years, with the population increasing from 1.09 to 1.35 million people from 1990 to 2020. The proportion of migrants and descendants increased from 8.3% to 23.5% over the total population in the same period [138]. Starting in the mid and late 1960s, Denmark has undergone multiple waves of migration originating from various countries such as the Nordic countries, Germany, the United Kingdom (UK), but also former Yugoslavia, Turkey, Morocco and Pakistan as part of workforce-migration program [64, 124, 138].

In the 1990s, there was an influx of migrants from Vietnam, Chile, Poland, Iran, Palestine and several African countries [64, 124]. These were followed by significant migration from Eastern European countries in the 2000s, following the enlargement of the EU, with migrants of Polish origin consisting the fourth largest migrant group in the city in 2020 [138].

Denmark's increasing migrant population has led to challenges regarding socio-spatial segregation, sparking a prolonged debate on the so-called '*ghettos*' [47] or '*parallel societies*' [83, 42], when the majority of the residents of an urban block are of non-EU origin and rely on government assistance [138]. According to a report from the Ministry of Immigration and Integration in 2019, the number of immigrants and their descendants in Denmark reached 793,601, accounting for 13.7% of the total population [95]. Among them, approximately 20% resided in the capital region, with the highest concentration of migrants found in Ishøj [40%; 28].

The bivariate map in Figure 3.4 illustrates the mean population density and the mean proportion of migrants in relation to the total population, aggregated from the grid cell level ($\mathrm{DST_g}$) in the parish boundaries. The deep blue and purple hues in the inner core of the city indicate a high population density and a significant presence of migrants – moderate and high proportions, respectively [53]. In the area of Tingbjerg, one of the identified parallel societies in Denmark, the average proportion of migrants exceeds 75% of the total population. Additionally, the outer peripheral zone, characterised by lighter shades of purple, shows a lower population density and a lower concentration of migrants. In these areas, the average density is 100 and 200 persons per grid cell, with migrants accounting for an average of 7-21% of the population [53].

The presence of diverse communities coupled with the ongoing discussions on segregation makes Copenhagen an ideal case study for our experiments. Approaches as the one explored in this study can contribute in the recognition of potentially problematic areas where preventive and small-scale actions can take place, as opposed to the suggested by the Danish State strategy of demolishing blocks [145, 151].

Copenhagen also offers advantages as a case study in terms of access to detailed data. Similar to the Netherlands, DST maintains a population register called the central person register (CPR), which is updated daily. This register collects information on international and internal mobility, deaths and births [29]. Access to the 100 m gridded dataset was only granted under restrictive conditions within a DST-controlled environment because potentially sensitive information could be included in sparsely populated areas. During the COVID-19 pandemic, obtaining such access proved to be a time-consuming process, and subsequent technological limitations made it difficult to run long and computationally demanding simulations.

Despite these limitations, this study investigated historical migration patterns and their relation to urban features and population distribution by education and income level (sec. 11). Using the openly available data at parish level, the relations between Danes, Western and non-Western migrants were also explored using compositional data [CoDa; 5, 6] techniques [37].

Papers A, B and C (sec. 8, 9, 10) focused on the case study of the capital region of Denmark, known as Greater Copenhagen. The study used a gridded dataset with a resolution of 100 m. In Paper D (sec. 11), the study aggregated the gridded dataset to a resolution of 1 km, which was used for the analysis (DST_{g1}).

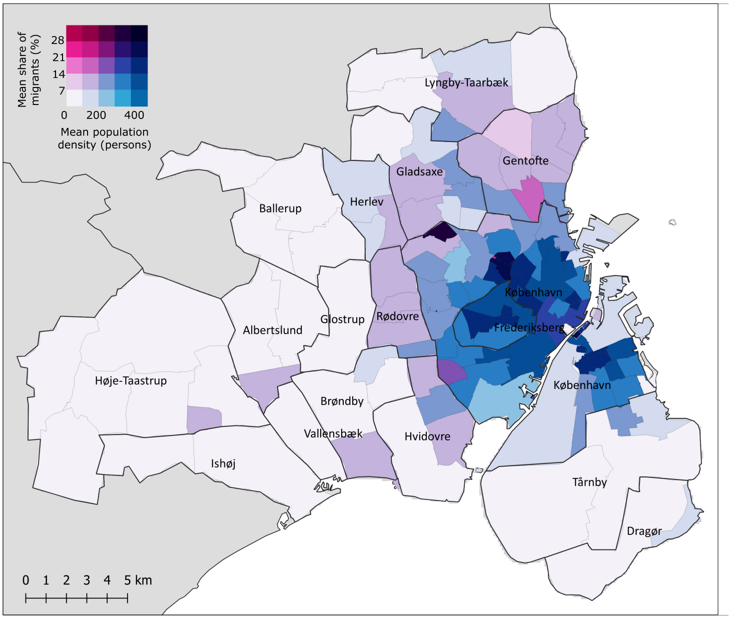


Fig. 3.4: Population density and share of migrants in Copenhagen in 2018. The area of Tingbjerg in dark blue stands out, where the average share of migrants per grid cell is over 75%. Source: Georgati and Keßler [53], modified by the author.

3.3 Kraków

Kraków presents a contrasting picture to the above cases. Before 2010, the scale of migration was insignificant – less than 5,500 people in total, mainly from Ukraine – and consisted of circular migrants who travelled regularly between the two countries. After the outbreak of the armed conflict in Crimea, many Ukrainians moved to Poland as economic migrants and students [34]. Thus, since 2015, Poland has transformed at an unprecedented pace from a '*migrant-exporting*' economy into a host and destination country for non-EU Europeans [139].

The rapidly changing image of Kraków is represented in the two bivariate map models in Figure 3.5. These models illustrate population changes in terms of individuals and migration shifts as a proportion of the total population over two-year intervals: 2013 to 2015 in the upper map and 2019 to 2021 in the lower map. Both maps highlight areas of significant divergences. Between 2013 and 2015, there was an exodus from the city centre, with the native and migrant populations in the coral areas decreasing and likely relocating to the surrounding less urban areas in light blue. In the second examined period, more intense mobility is observed, with the migrant share and total population increasing in Grzegórzki and Dębniki, and decreasing in Bieńczyce and Stare Miasto.

Kraków's recent transformation into a destination city, attracting mainly third-country nationals to the EU, makes it an interesting case study for examining patterns of migrant distribution in the early stages of their formation. As its ethnic diversity continues to grow, Kraków is expected to face similar challenges to Western European countries in terms of integrating migrants and promoting social cohesion [25]. These challenges include issues related to residential segregation and the concentration of migrants in certain areas.

The very small foreign population in Kraków before 2010 was dispersed throughout the city, with slightly higher concentrations in Stare Miasto and Krowodrza. From 2010 onwards, migrants began to move to districts such as Stare Miasto, Grzegórzki, Krowodrza and Prądnik Czerwony, and to a lesser extent to Bieńczyce, Podgórze Duchackie, Mistrzejowice and Bronowice. Nowa Huta, an economically marginalised district with a bad reputation, but lower rents and good public transport, saw increased migration activity [116]. From 2015, the central districts of the city attracted migrants, as did the peripheral districts of Mistrzejowice and Bieńczyce, where there was a supply of rented accommodation and new blocks of flats accommodated mainly foreigners [138]. The outcome of this mobility in 2019 is shown in Figure 3.6, which maps the average share of migrants at the census level and the average population density aggregated from the grid cell level to the statistical regions.

In contrast to the previous cases, the collection of population data was a major challenge due to the lack of a central register that is updated based

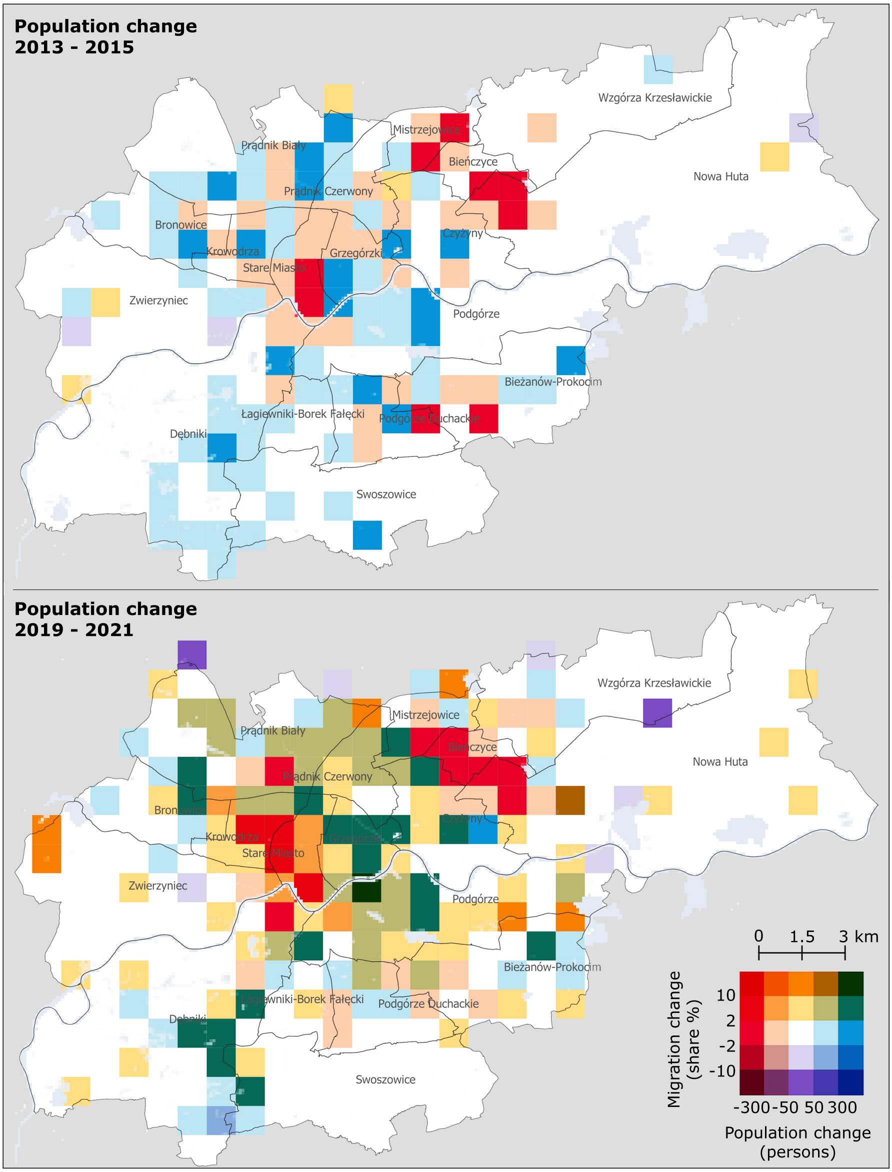


Fig. 3.5: Estimated change in the distribution of migrants in Kraków from 2013 to 2021 in relation to the corresponding change in the total population. The change is shown in two time steps, between 2013 and 2015 (top) and 2019 and 2021 (bottom). White indicates minimal change in both the total population and migrant share; red indicates increasing migrant share with decreasing total population; green indicates increasing migrant share with increasing total population; and dark red and blue indicate decreasing migrant share with corresponding decrease or increase in the total population.

on population movements. Due to this lack, a unique dataset in Poland was developed through the cooperation of the FUME team with the Municipality of Kraków and the Lesser Poland (Małopolskie) Voivodship Office. The Polish gridded dataset includes several characteristics of migrants residing in the city using register data from these institutions. The dataset combines data on migrants from the register of people legalising stay on the territory of Poland and the register of inhabitants of Kraków [139].

This dataset was used to analyse residential segregation in Kraków in 2019 using the individualised neighbourhood approach by Stonawski *et al.* [139] – contribution not included in this dissertation. The paper reveals a relatively low level of migrant segregation compared to cities with long history of migration in Western Europe. Although migrants constitute 4.2% of the population, half of the residents live where foreigners constitute less than 2.2% of their 200 closest neighbours. Around 45% of migrants would have to relocate to achieve a more balanced distribution, with certain districts showing higher concentrations of Southeast and South Asian, and North and Latin American migrants. Migrants from Ukraine and other Eastern European countries are more evenly dispersed. Notably, foreign residents in Kraków are well-educated and live in central and prestigious residential areas, in contrast to Western European trends where migrants have lower levels of education and settle in less desirable by the natives locations [139].

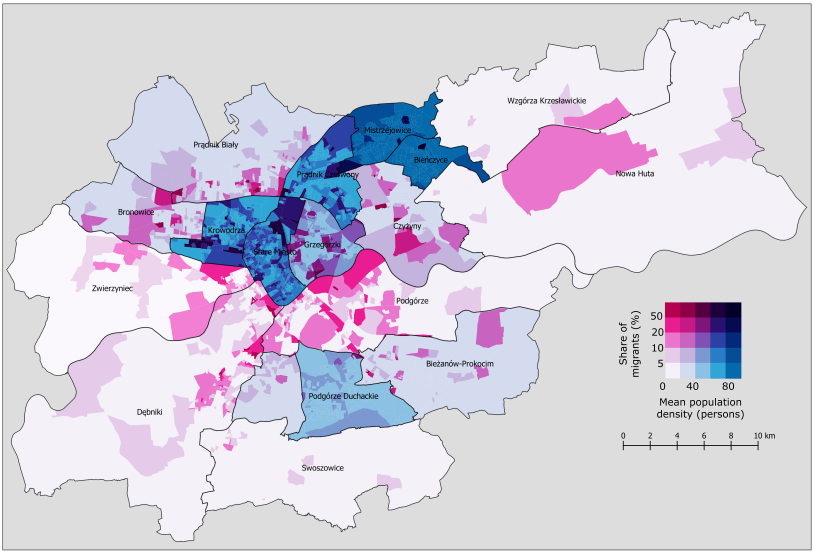


Fig. 3.6: Average share of migrants and average population density in Kraków in 2019 at the boundaries of statistical regions and census enumeration areas. Source: Stonawski *et al.* [139], modified by the author.

3.4 Rome

After a long period of emigration, Italy and Rome began to become a major migration centre in the 1970s. The foreign population in the country has grown from 350,000 in the early 1990s to over 5 million today. Rome is particularly attractive to migrants because of its role as the capital of Italy, home to diplomatic missions and the Vatican City, which promotes and coordinates support and assistance for migrants. The city's rich cultural heritage and the presence of universities and international organisations attract highly skilled migrants but most migrants in Rome work in low-skilled jobs, primarily in the service, tourism, construction and domestic sectors [138].

The urban core of Rome depopulated between 1970 and 2000, with the population sprawling to the suburbs for cheaper housing. However, the economic recession of 2008 and 2011 with the consequent decrease of real estate prices led to the re-urbanisation of its urban centre. From 2000, the number of foreigners has been increasing rapidly and in 2015 exceeded the 300,000 persons, following the regularisation of many migrant workers and the entry of Roumania into the EU in 2007. Despite the economic crisis, the number of entering foreigners continued growing until 2015 when restrictive national migration policies were implemented and socio-economic imbalances strengthened.

During the study, collecting gridded historical data on migrant distribution in Rome posed challenges. However, a structured dataset of estimates was successfully compiled based on census tracts, focusing on migrants from different continents and major countries. For 2020, data was collected at grid cells for validation purposes, but values below 4 were obscured due to confidentiality concerns.

Another data-related issue we encountered was over- and underestimations of the populations in the census tracts data. The population register of 2014 revealed that approximately 25% of the foreign population in Rome had not been accounted for in previous census operations. Furthermore, the problem of '*fake addresses*' arose, leading to unrealistically high concentrations of migrants in central locations. This occurred because institutions assisting migrants in overcoming administrative obstacles allowed them to register with a fake address as their place of residence for the municipality. This approach enabled homeless individuals or those without a fixed address to obtain identity cards, social/health assistance and official communications from the municipality.

Figure 3.7 and 3.8 display the distribution of the population and the share of migrants at census tracts (CNR) and gridded ($\mathrm{CNR_g}$) level in 2020. Figure 3.7 offers a comprehensive view of the entire municipality, while Figure 3.8 provides a closer examination of specific regions within the city (Map A, A1, A2), namely the city centre and its urban periphery. This simultaneous presentation of aggregated and gridded population distributions underscores the significance of local estimations in accurately capturing the density variations across different

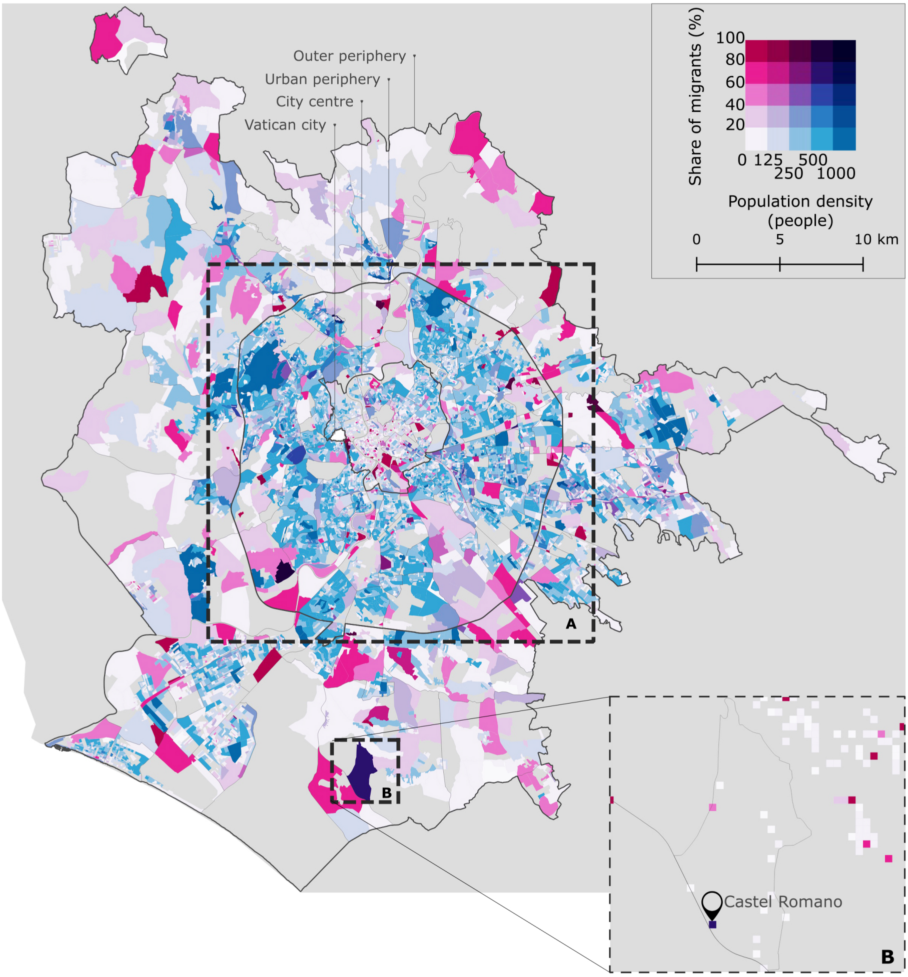


Fig. 3.7: Population density and share of migrants in Rome at the level of census tracts in 2020. The key map shows the three sub-areas and the 155 urban areas of the Municipality of Rome, as well as the location of the two zoom-in areas: Map A focuses on the urban core of Rome (city centre and urban periphery) and is shown in Figure 3.8 at the resolution of 100 m; Map B zooms on the southern part of the outer periphery, in the urban area where the Castel Romano Roma camp is located.

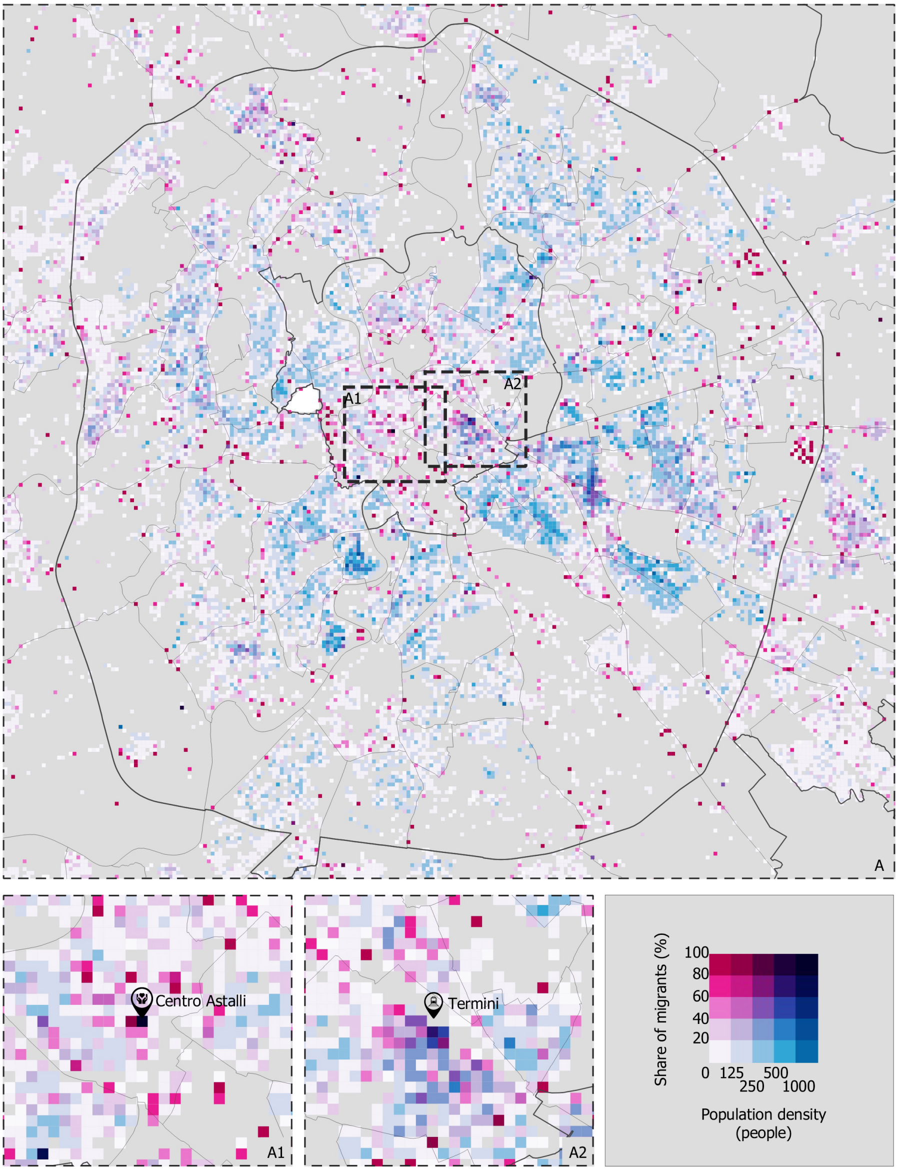


Fig. 3.8: Population density and share of migrants in Rome in 2020 at a resolution of 100 m. Map A1 focuses on Piazza Venezia and Centro Astalli – an institution that helps migrants and refugees – and Map A2 on the central train station (Termini), where the high concentration of migrants is linked to the active retail trade. The key map is shown in Figure 3.7.

locations.

For instance, Map A2 in Figure 3.8, which focuses on the area around the central railway station (Termini), shows the varying densities among urban blocks that generally exhibit a high presence of migrants (i.e. Chinese, Bangladeshis) engaged in active retail businesses. Map A1 is centred around Piazza Venezia and its surroundings, which are not primarily residential zones. This area hosts an institution that helps migrants and refugees – Centro Astalli at via degli Astalli 1 (https://centroastalli.it/). Despite the efforts made in recent years to address the issue of '*fake addresses*', with the Municipality of Rome instituting official virtual addresses for the homeless, alongside specific census tracks, the gridded dataset reveals high population density and migrant concentration in this particular location, which does not conform to the character of the surrounding area.

Lastly, Map B in Figure 3.7 zooms in on a census tract that is presumed to be densely populated, with a high concentration of migrants. In fact, it represents a low-density suburban area that hosts Castel Romano – one of the largest Roma camps in Rome [89] – on its southern borders. This serves as a striking example of the complexities and nuances inherent in the spatial distribution of populations and migrant communities from datasets aggregated at various administrative levels.

The collected at census tracts level dataset for Rome was disaggregated at grid cell level through areal weighted interpolation due to the small size of the census tracts in central areas of the city. Then this dataset was used for initialising the future projections. It is expected to be used for further analysis of the spatio-temporal patterns of migration in Rome in a working paper.

4

Summaries of articles

Summaries of articles

The dissertation consists of three articles of which the candidate is the first author and one article of which the candidate is the second author. They are submitted for consideration together with the dissertation itself. This section consists of summaries of these core articles (Papers A, B, C and D in sections 8, 9, 10, 11 respectively). The aim of each paper in the context of the PhD dissertation is described, followed by an overview of the incentives behind it, the proposed methods, the main findings and the discussion points. Although the FUME project included four case studies, only two of them were examined in these papers due to the quality and availability of data at an early stage of the research. Findings on all case studies and the implemented methodology for connecting the regional projections to the local distributions follow in section 5.

4.1 Spatial disaggregation of population subgroups leveraging self-trained multi-output gradient boosting regression trees

Georgati, M., Monteiro, J., Martins, B., & Keßler, C., published in the 25th AGILE Conference on Geographic Information Science: Artificial Intelligence in the service of Geospatial Technologies (Vol. 3, pp. 1-14), in 2022

The aim of this paper within the PhD dissertation was to explore the potential of multi-output ML models for simultaneously disaggregating population cohorts in comparison to classic dasymetric approaches and single-output ML models.

This work was motivated by the importance of accurate and consistent estimates of the distribution of population cohorts at fine spatial resolution. Such estimates are necessary to support public administration functions in various areas, including urban development, environment and the economy. The paper addressed challenges related to differences in sampling regimes, sample sizes and data collection methods across different time periods and geographical regions, as well as privacy concerns related to high-resolution population datasets through spatial disaggregation.

Specifically, it introduced a novel method using self-training regression models and innovated over previous studies by simultaneously downscaling counts for multiple population cohorts through the use of multi-output ML models. The methodology involved estimating weights based on classic dasymetric approaches such as simple pycnophylactic or dasymetric interpolation, using the GHS-POP layer [129]. These weights were the labelled dataset and were used to initiate the training of the regression model, while the produced disaggregated dataset from each iteration was adjusted to maintain the original source zone counts. This adjusted disaggregated dataset was then used as the labelled dataset to train the subsequent iteration.

The authors conducted experiments with multi-output models using the RF and GB algorithms and compared them with the two above-mentioned traditional dasymetric approaches. The multi-output GB model, apart from its advantages in learning by training from previous iterations, advanced by using a task-specific loss function that combined the interrelated variables. Specifically, the authors combined the MAE – which estimated the error in the predictions of each individual population subgroup – with the RMSE – which added the penalties for errors in the sum of the subgroups of each class.

The experiments were conducted on two case studies, focusing on the

Municipality of Amsterdam and the region of Greater Copenhagen. The study examined classes based on age and migration background, which were further divided into subgroups. The selected case studies differed from each other in several aspects (i.e. extent, source zone, number of classes and subgroups). In Amsterdam, the area extended over one single municipality that comprised multiple neighbourhoods. In Copenhagen, the case study area encompassed 17 municipalities. These variations enable the application to be generalised to other cities or groups based on the respective requirements.

The study used a two-level experimental design to examine the suitable combinations of training data and tree depth. The authors analysed data from the case studies and found that the optimal combination of ancillary data varied between the cases due to variations in data availability. However, they also reported the corresponding results for Copenhagen, where they could use the best combination of datasets from the Dutch case.

The results were evaluated using ground truth data at the target resolution provided by the national registers. Based on the error metrics, the single-output RF and the multi-output GB regressor with 500 trees outperformed the other experiments in both cities. The evaluation of the results also included their visual inspection, where the mapped predictions of the models over the ground truth data were examined. Finally, the absolute percentage error of the two best models was visualised, showing the high performance of the multi-output GB model in densely populated areas.

Overall, the paper presented a spatial disaggregation approach that incorporated self-training, regression models and multi-output techniques. The authors validated the method through two case studies and highlighted its advantages for predicting multiple interrelated variables in densely populated urban areas. The results showed that the approach captured the spatial heterogeneity and dependencies between factors, breaking the smooth distributions of the seminal disaggregation algorithms. However, further analysis was suggested to address the difficulties in interpreting the results due to the small size of the grid cells, the large extent of the study area and the pixel-level evaluation. The authors also recommended further research into other libraries or models that would reduce the long training time required by the GB trees, and the exploration of other sources of training information.

4.2 Modelling population distribution: a visual and quantitative analysis of gradient boosting and deep learning models for multi-output spatial disaggregation

Georgati, M., Monteiro, J., Martins, B., Keßler, C., & Hansen, H. S, submitted to Transactions in GIS in June 2023

The aim of this paper within the PhD dissertation was twofold: firstly, it extended the research presented in Paper A to deep learning modelling in order to improve prediction accuracy. Secondly, it employed several assessment methods to evaluate the error and its spatial distribution, thus contributing to the challenge of developing more generalisable spatial models.

This paper built upon the previous study, Paper A, to address some of the challenges encountered. These challenges included the long training of GB tree models and difficulties in interpreting the results due to small size of the grid cells, the large extent of the study area and the pixel-level evaluation. The authors contributed to these challenges by first validating and refining the methodological approach and then conducting an in-depth quality assessment with alternative options.

In particular, the authors relied on the methodology of Paper A and expanded the regression modelling from ensemble learning to fully convolutional neural networks. They incorporated the task-specific loss function to produce disaggregated estimates of interconnected population groups comparable to those obtained from GB trees. Experiments with ideas borrowed from the modelling approach of Monteiro *et al.* [98] were examined to improve the quality of the output in terms of spatial variability. Models produced by classic dasymetric approaches and GB trees, as presented by Georgati *et al.* [54], were used as baselines.

Experiments were conducted in the region of Greater Copenhagen using detailed input datasets on building characteristics to train the models. Multiple demographic variables, divided by age and migration background, were simultaneously downscaled from the municipal level on high resolution grids (100 × 100 m).

In terms of quality assessment, this paper presented an in-depth evaluation of the spatial distribution of the error. The evaluation was again performed against ground truth data at the target resolution and provided a direct and accurate assessment of the quality at pixel level. However, the pixel-level evaluation did not provide conclusive results. Particularly in densely populated areas the small cell size relative to the study area posed challenges. Therefore, the authors proposed an accuracy assessment approach using individualised neighbourhoods [133] to assess the error among groups of cells within short distances.

Georgati *et al.* [52] reported results with error metrics such as the MAE, but also interpreted the error distribution by land use and in relation to the underlying patterns of the training data and the urban development of the city. Utilising the concept of individualised neighbourhoods, insights into the spatial accuracy of the models were gained by averaging the error across zones. The authors elaborated on focus areas and chose to analyse two migrant groups whose residential distribution patterns differed. They showed that by using different types of quality assessment, valuable interpretations could be achieved, localising the error and the reasons behind it.

The authors discussed the advantages of a multi-output approach over computing individual variables separately, especially when the variables were interconnected, and improved predictive accuracy. They also highlighted the contrasting features of the two examined architectures, with the deep learning approach improving performance and prediction accuracy at the expense of a more complex architecture requiring time-consuming tuning and refinements. The quality of the initial estimates and their impact on the performance of all modelling approaches in predicting the distribution of specific migrant groups, particularly in regions where the majority of the population was foreign, was discussed. Lastly, the authors underlined the limitations of utilising training datasets of building features in cities lacking detailed data and the sensitivity of the deep learning approach to such datasets.

4.3 Random forest variable importance measures for spatial dynamics: case studies from urban demography

Georgati, M., Hansen, H.S., & Keßler, C., submitted to the International Journal of Geo-Information in August 2023

The aim of this paper within the PhD dissertation was to investigate the determinants of residential distributions in a data-driven study focusing on the significance of the ethnic composition of the area and the variations among residents of different migration backgrounds.

The identification of the factors that drive residential choices and influence the distribution of different demographic groups in cities has been a significant topic of discussion throughout the project. Providing such an agenda would aid in modelling to understand the reasons why an area is more attractive to certain migrant groups, and would more effectively guide the selection of training data for the above work. This paper proposed a data-driven approach to rank and interpret the importance of features in migrant residential location modelling, unlike previous studies that have primarily focused on qualitatively analysing these determining factors.

Initially, the authors collected various determinants of residential distributions, including demographic and socio-economic factors, by reviewing the existing literature on residential modelling and choices. They used high-resolution grids to spatially represent these factors and register data on the distribution of migrants at the corresponding resolution. To rank the importance of the factors and training datasets, the authors employed variable importance measures [VIMs; 60, 26], a standard technique to extract interpretable information on the contributions of training variables from RF.

In this case, the authors built a simple supervised learning framework to explain a complicated phenomenon. The dataset of the distribution of each migrant group was used as a label, whereas data on demographic, building and neighbourhood attributes, together with the distribution of other migrant groups, were used as dependent variables. The analysis was carried out for the reference year 2018 and in two phases; first for groups of aggregated RoOs and then for individual CoOs in Amsterdam and Copenhagen. The findings from the aggregated groups highlighted the key factors associated with each migrant group. The findings from the groups by CoO were related to the size of the foreign population. Such a relation was considered important to investigate in order to understand whether the attraction was linked to pre-existing centres and facilities or cultural bonds between ethnic groups.

Interpretations of the findings were drawn upon the influences of ethnic groups. High importance scores were observed between migrants from the former colonies and the Middle East and Africa in Amsterdam, and between Western EU and other Western countries in Copenhagen. Associated factors were the type of housing – which may be related to underlying patterns of economic status of different groups – and distribution by age – which may be related to the period during which particular migrant groups settled.

The authors also examined the relationship between the distribution of migrants and house prices in 2008 and 2018. The study used statistical analysis to detail the variation in house prices across different migrant groups between the studied years. Bivariate maps were used to illustrate the high-price regions in relation to the distribution of migrants from Middle Eastern and African and Western countries as a proportion of the total population.

In the first part, the study captured a specific moment in time and focused on the spatial distribution of migrants at that particular point in time. This methodology allowed for an in-depth understanding of the underlying factors and their urban dynamics for each group. The authors identified migrant groups from neighbouring CoOs that received high importance scores to each other. For example, they pointed to Balkan migrants in Copenhagen and African and Latin American migrants in Amsterdam. Taking both parts of the analysis into consideration, the authors found evidence of segregation in both cases, with Western migrants, originating from within the EU and outside it, appearing clustered together in areas with high-quality housing and low concentrations of non-Western migrants.

This paper highlighted the importance of the neighbourhood's ethnic composition in the residential distribution of people from different migrant backgrounds at the fine geographical scale of 100 m grid cells. It contributed to the assessment of the impact of cultural background on the residential locations of migrants by examining a large number of CoOs that had not previously been studied. The authors ranked the migrant groups identified by the ML-based model as most influential on the distributions of other minorities and evaluated the relationship between ethnicity, cultural and geographical proximity of CoOs and residential distribution in the destination cities. They suggest further research on additional time steps for a comprehensive evaluation of the underlying factors and for relating the results to theoretical frameworks around integration processes.

4.4 Analytics of historical human migration patterns: use cases of Amsterdam and Copenhagen

Kveladze, I., Georgati, M., Keßler, C., & Hansen, H. S, published in the Journal of Location Based Services in 2023

The aim of this paper within the PhD dissertation was to explore patterns in the distribution of different migrant groups in relation to the native population and other socio-economic and urban characteristics using different visualisation techniques.

This study was inspired by the need for comprehensive visualisation methods to explore and analyse human migration dynamics. Visualising rich and complex spatial datasets from population registers can be challenging in terms of scale, heterogeneity, spatial and temporal resolution of the data and, most importantly, communication and interpretation to extract meaningful insights. This paper suggested mapping gridded population datasets, that preserved data attributes, and the representation of statistical computations. This approach enhanced visualisation and facilitated in-depth analytics of migration patterns.

The authors investigated the correlations between different data attributes using a single symbol visualisation mechanism; they employed univariate and bivariate choropleth mapping, a proportional symbol method and a combination of these techniques through multiple blending modes. The visualisation approach ensured consistent and efficient statistical analysis, allowing the exploration of changes, identification of patterns, relationships and trends using additional quantitative information present within the grid cells.

Bivariate maps with diverging colour schemes displayed positive and negative ratios between two attributes (i.e. population and migration change between two periods) with darker hues at either end from the centre of the data range. Complex thematic representations with bivariate colour maps and sequential colour schemes simultaneously displayed two variables with high and low values, while also incorporating proportional size symbols for additional data characteristics. Such visualisations enabled the comparison between two or more variables, reducing cognitive workload and facilitating understanding of complex relationships within the dataset for attribute value distribution in urban areas.

One of the main priorities of the research was to achieve visually appealing results while also helping the map reader to orient themselves in the geographical space beneath the grid cells. This approach optimised the visual impact by enhancing the contrast of darker base map areas and reducing the visual impact of lighter areas. Despite the complexity of the map design, the produced models enhanced the analytical capability of the information displayed and revealed underlying correlations between data characteristics.

The visualisations uncovered complicated relationships between data attributes across the case studies (i.e. Amsterdam and Copenhagen) in the most effective and efficient manner. The analysis was performed on 1 km grids and focused on the distribution patterns among native and migrant populations and their changes between 2010 and 2018.

The produced visualisations were informative and engaging, showing dynamic changes in migration over time. They showed an uneven distribution between non-EU and EU migrants in different urban areas, which could be explained by the presence of diaspora communities as a pull factor. The maps also presented a stronger representation of non-EU than EU migrants in the majority of the grid cells, with a few exceptions in Copenhagen. Comparing the change between 2010 and 2018 in the district of Zuidoost in Amsterdam, the maps showed a decrease in the number of non-EU residents, as mentioned in section 3.1. The research also highlighted the ‘moving in and out’ activities of the local population, with migrants being slightly more active in Amsterdam than in Copenhagen. Finally, the attractiveness of certain central areas was underlined, possibly because of the denser presence of cultural and educational facilities.

Overall, the study provided valuable insights into migration patterns between the native and migrant populations in the examined cities. However, the potential loss of information due to the aggregated level of data was acknowledged. The authors suggested further investigation on this topic to confirm patterns related to amenities, along with the inclusion of other demographic attributes such as age and gender. These factors could greatly assist in understanding the underlying connections between different demographic groups within the urban fabric and provide valuable insights into the specific factors influencing the decision-making process.

5

Gridded migrant estimations for FUME

Gridded migrant estimations for FUME

Following the presentation of the methodological journey in the spatial disaggregation of multiple demographic groups and the exploration of their patterns, this section reviews the FUME requirements, mentions methodological adjustments and provides a brief overview of the outputs generated for FUME for the four case studies.

Due to time constraints, local projections were carried out and presented for two out of the six developed by FUME scenario projections: the benchmark and the baseline scenarios. The benchmark scenario reflected the SSP2, incorporating the global impact of the COVID-19 pandemic. It also incorporated observed trends in internal and international migration between 2015 and 2019 at the regional level of each country. In contrast, the baseline scenario assumed no international migration and zero net internal migration at the regional level. Although such a scenario is unrealistic in practice, it provided a reference point for analysis and testing, allowing the understanding of the impact of migration on population dynamics under simplified conditions. Selected visualisations of the historical (2020) total and migrant populations and projected migrant distributions in 2030 under the benchmark and baseline scenarios are presented on the following pages (Fig. 5.1 - 5.16).

As mentioned above, the adopted disaggregation approach used self-training, where weights were used to estimate a 'teacher' at the target resolution to initiate the training of the ML model. In the experiments of the aforementioned papers, these initial estimates were generated for each group through dasymetric interpolation based on the GHS-POP layer, which represented the distribution of the total population. Nevertheless, discrepancies and deviations were identified in the evaluation of the results for specific migrant groups, especially in regions with a significant population of foreign origin, as discussed in Paper B. These inconsistencies were attributed to the strong influence of the total population's proportional distribution during the dasymetric interpolation. Consequently, additional measures were required to differentiate the distributions of the various migrant groups and address these inconsistencies.

To ensure that the results aligned better with historical patterns, methodological adjustments were implemented, thereby enhancing the quality of the generated results. The GHS-POP layer was replaced by migrant-specific layers at their most recent available distributions, which provided an indication of the currently more attractive regions for each group. Moreover, instead of disaggregating the total population of each group, the change that occurred between the latest available historical data and the projected year was used. The generated distribution of the change was then added to the latest historical distribution as the final product. These adjustments ensured alignment between the current spatial distribution of migrants and the projected distribution, and provided greater flexibility in capturing variations and dynamics within population groups.

To explore the potential applications of such an approach, the output produced was used to analyse migration and residential segregation in the four cities in combination with the historical data collected. This report [138] presented a comprehensive description of residential allocation for each city, accompanied by a mapping of the distribution of the corresponding groups over different time periods. It also explored possible futures of the spatial allocation of the groups studied in these cities, based on the two scenario projections.

Historical and projected data from the two scenarios were analysed using a range of metrics at both the grid cell and neighbourhood level. Segregation was analysed as the evenness of the distribution of two groups across neighbourhoods using the dissimilarity index [92, 90] – a common measure for understanding segregation and the representation of a group in a neighbourhood [138]. A measure of the likelihood of a foreigner meeting someone from the same migrant background in their immediate neighbourhood was also used to measure the residential segregation experienced by different migrant communities [138].

A key finding of the report was that despite considerable differences in migration history and in the structure and size of the foreign population, residential segregation measured at the grid cell level remained consistent in Amsterdam, Copenhagen and Rome. In contrast, Kraków, with a smaller migrant population and a recent influx of migrants, showed a relatively uniform distribution of migrants across the city and a declining dissimilarity index. The disaggregated projections suggested that national and regional migration trajectories had a relatively moderate impact on residential segregation within cities [138]. Extreme scenarios, such as the baseline and benchmark scenarios, did not significantly affect long-term residential segregation patterns. However, it is important to recognise that the model cannot capture all the factors that influence the residential distribution of migrants, and training the model on recent data where some migrant groups are minimally represented could result in overly uniform local predictions of their distribution. While this analysis showed a trend towards decreasing dissimilarity in the explored cases, it is also reasonable to assume that as small migrant communities expand, they may cluster in certain neighbourhoods, potentially increasing segregation [138].

Although the discussion of the obtained results is beyond the scope of this dissertation, the use of gridded data showed clear advantages over the use of administrative unit aggregated data, effectively avoiding the modifiable aerial unit problem [MAUP; 111, 112]. The MAUP prevents accurate comparisons of levels and patterns of segregation between areas, particularly if they differ in size, because the units used to measure segregation affect the results of a segregation analysis [101, 8]. This approach allowed the authors to make comparisons across cities and to create uniformly sized neighbourhoods around each city resident, allowing for a comprehensive analysis from multiple perspectives and enriching the study at the grid cell level.

Amsterdam

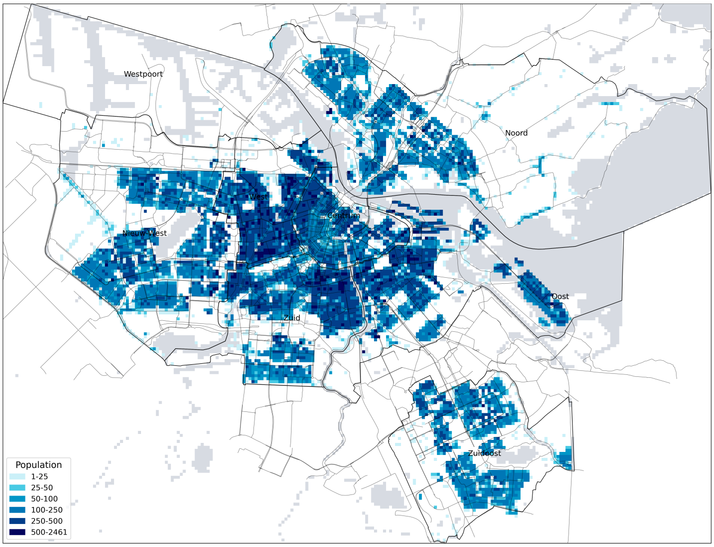


Fig. 5.1: Distribution of total population in Amsterdam in 2020 at 100 m resolution.

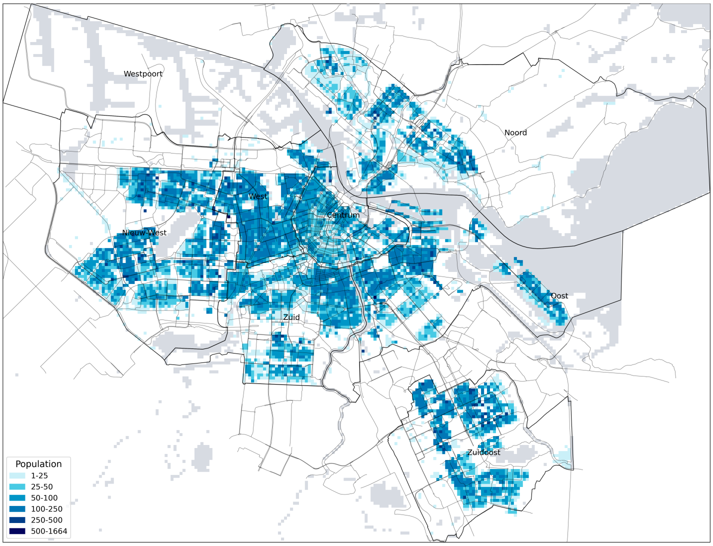


Fig. 5.2: Distribution of migrant population in Amsterdam in 2020 at 100 m resolution.

Amsterdam

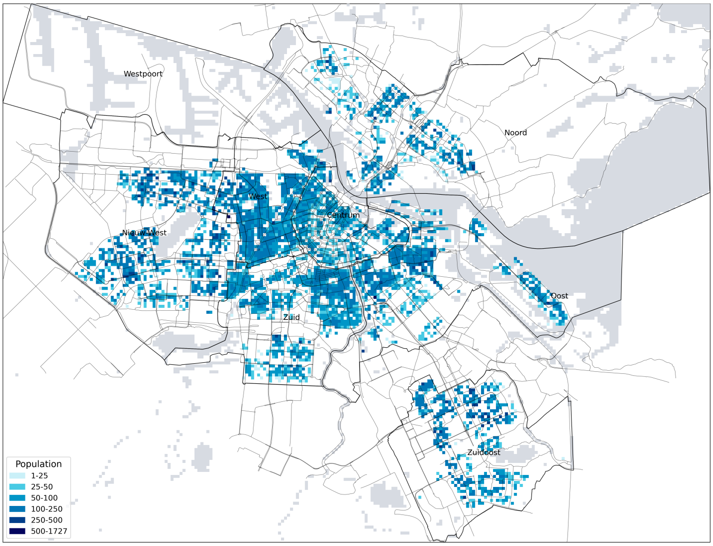


Fig. 5.3: Estimated distribution of the migrant population in Amsterdam in 2030 at 100 m resolution under the benchmark scenario.

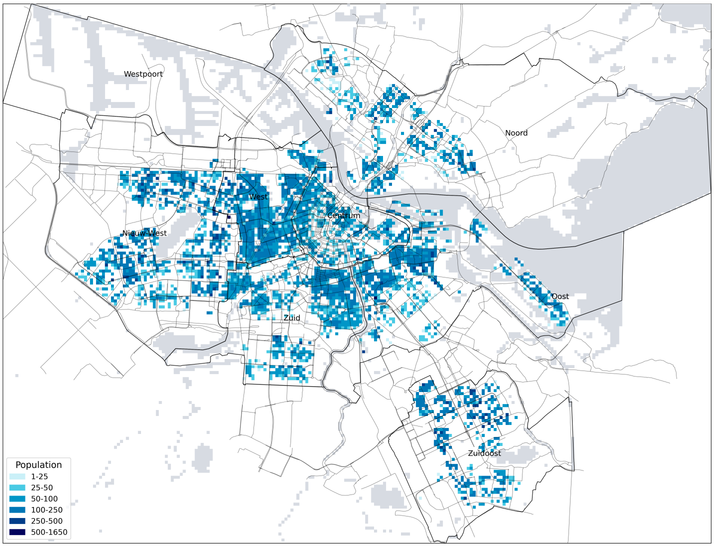


Fig. 5.4: Estimated distribution of migrant population in Amsterdam in 2030 at 100 m resolution under the baseline scenario.

Copenhagen

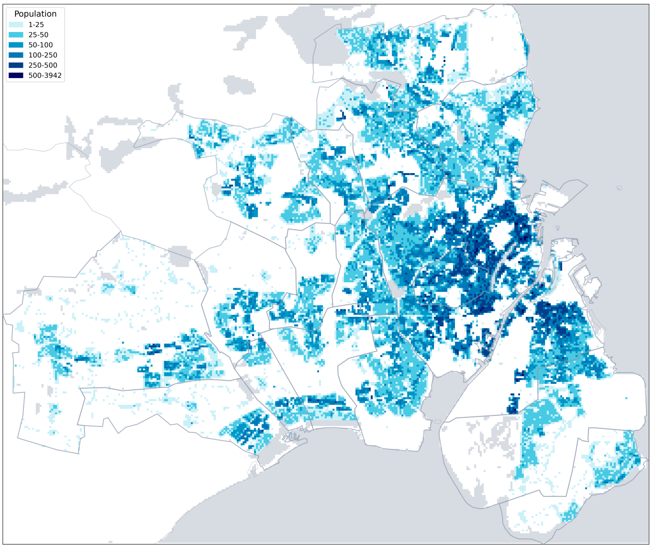


Fig. 5.5: Distribution of total population in Copenhagen in 2020 at 100 m resolution. Source: Sánchez Gassen [144].

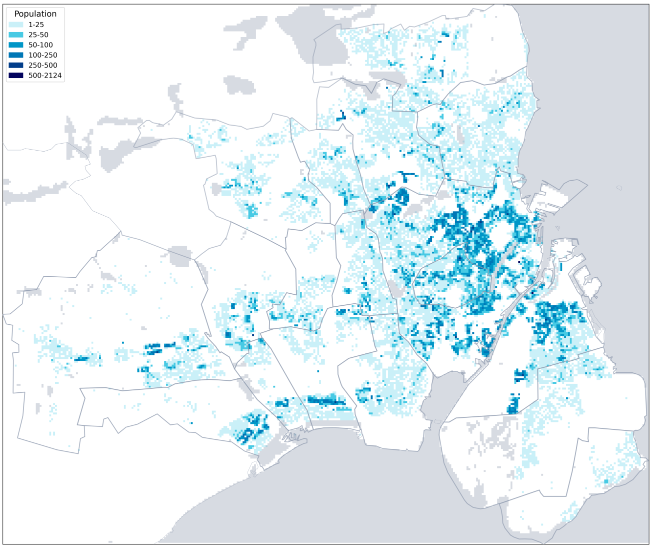


Fig. 5.6: Distribution of migrant population in Copenhagen in 2020 at 100 m resolution.

Copenhagen

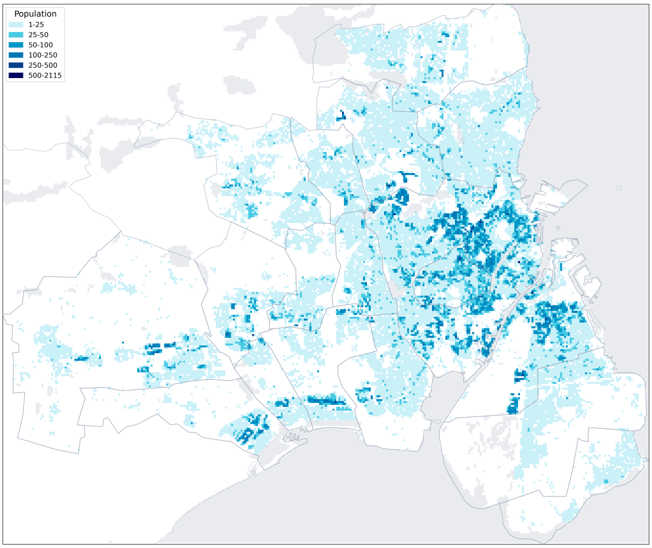


Fig. 5.7: Estimated distribution of migrant population in 2030 at 100 m resolution under the benchmark scenario.

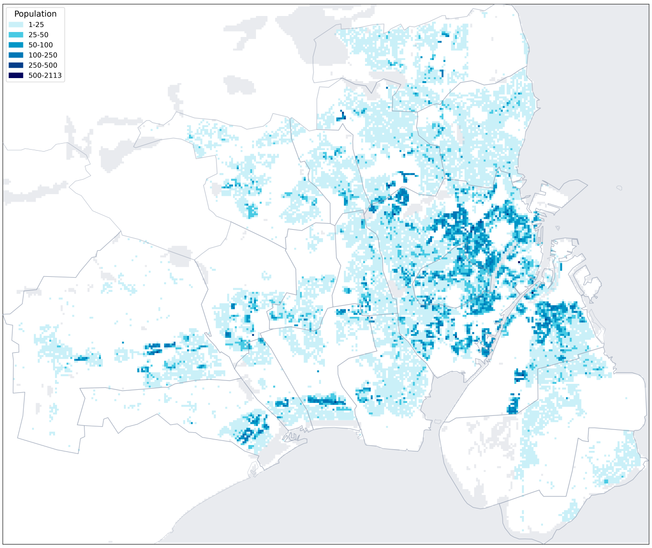


Fig. 5.8: Estimated distribution of migrant population in 2030 at 100 m resolution under the baseline scenario.

Kraków

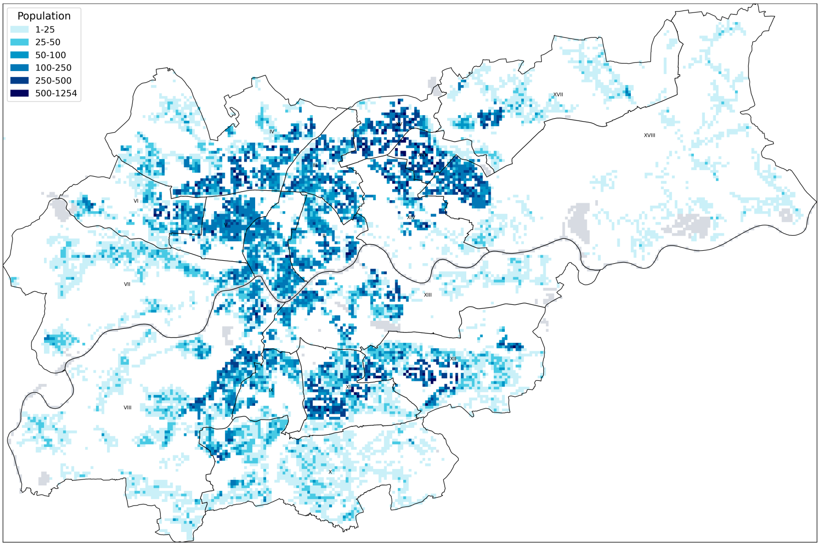


Fig. 5.9: Distribution of total population in Kraków in 2020 at 100 m resolution.

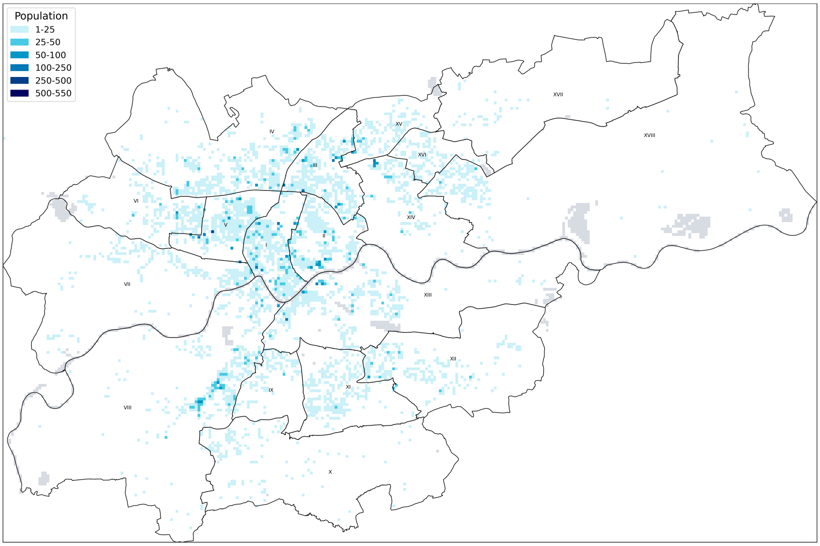


Fig. 5.10: Distribution of migrant population in Kraków in 2020 at 100 m resolution.

Kraków

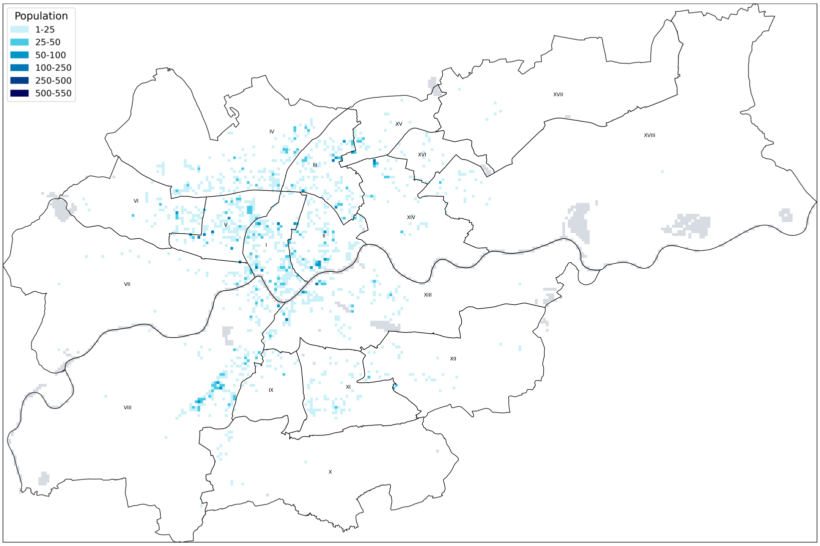


Fig. 5.11: Estimated distribution of migrant population in Kraków in 2030 at 100 m resolution under the benchmark scenario.

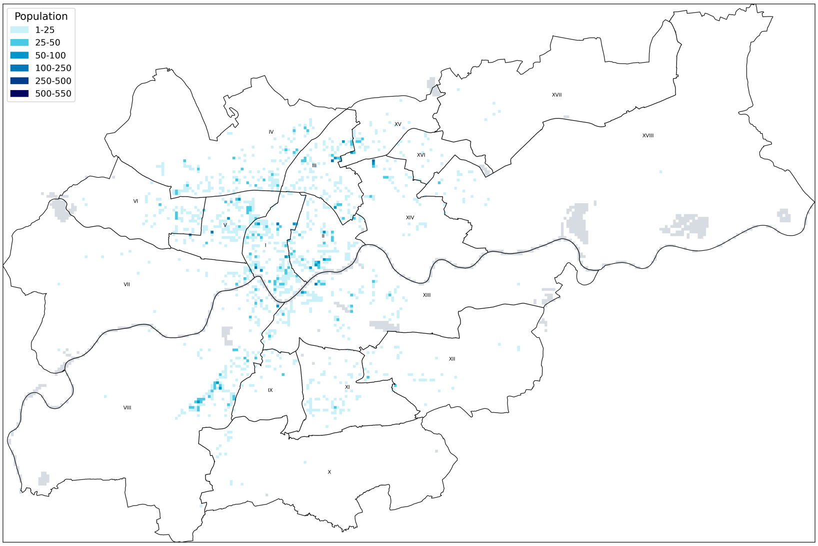


Fig. 5.12: Estimated distribution of migrant population in Kraków in 2030 at 100 m resolution under the baseline scenario.

Rome

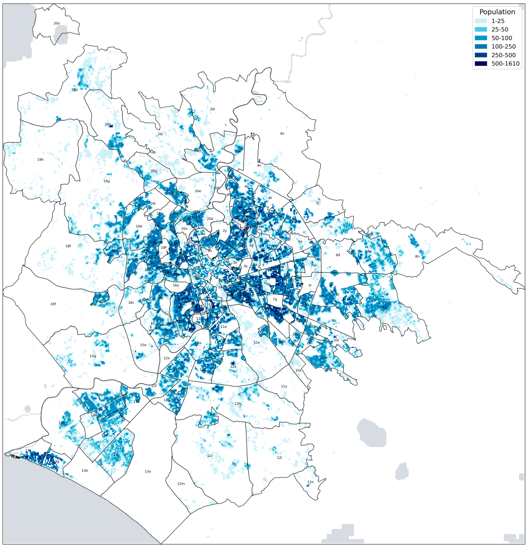


Fig. 5.13: Distribution of total population in Rome in 2020 at 100 m resolution.

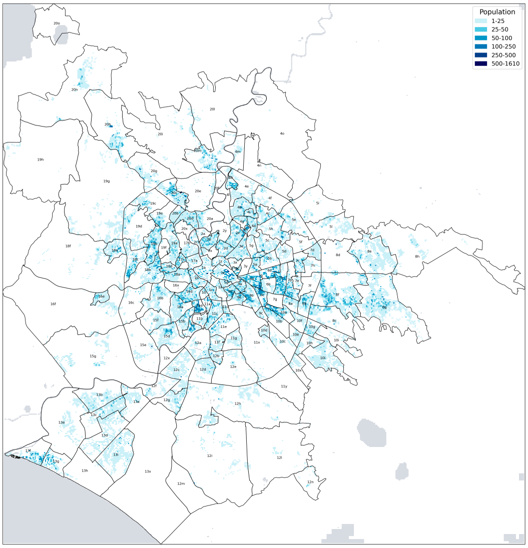


Fig. 5.14: Distribution of migrant population in Rome in 2020 at 100 m resolution.

Rome

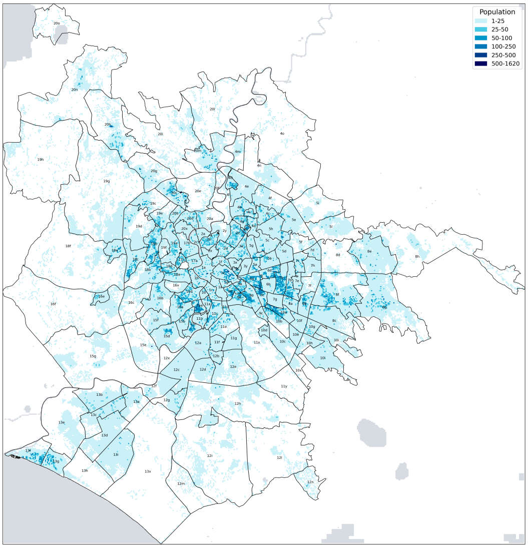


Fig. 5.15: Estimated distribution of migrant population in Rome in 2030 at 100 m resolution under the benchmark scenario.

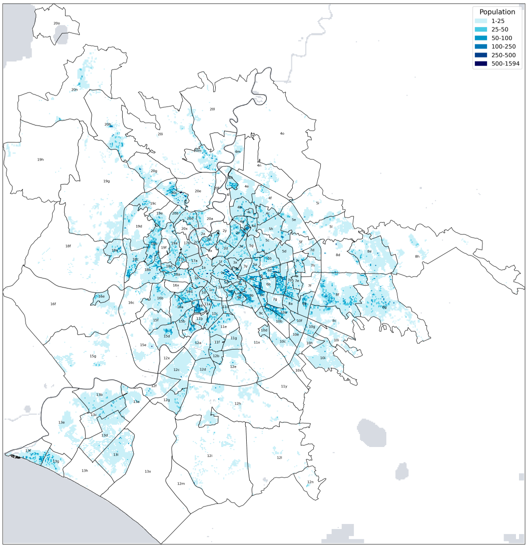


Fig. 5.16: Estimated distribution of migrant population in Rome in 2030 at 100 m resolution under the baseline scenario.

6

Discussion

Chapter 6.

Discussion

This PhD project was motivated by the increasing impact of migration on the urban environment of European destination cities. As part of a larger interdisciplinary project investigating migration and following an approach at multiple geographical scales, this work focused on establishing a connection between regional population group estimates and their local distribution at fine-grained resolution.

To achieve this connection, a spatial disaggregation methodology was employed that used ML to link aggregated population data to the specific residential locations of migrants in destination cities. The approach highlighted the potential of integrating ML into a disaggregation methodology for downscaling multiple demographic groups together, rather than one attribute at a time. The study also contributed to the discussion about methods that support the evaluation of the disaggregated products and their explanation in terms of model architecture and training variables. A detailed visual and quantitative analysis of the errors provided useful feedback for interpreting the results of spatial ML models.

Furthermore, the study contributed to the understanding of the factors that influence the residential distribution of migrants. Following a data-driven approach, it related migrant distribution to features such as the ethnic and demographic composition of a neighbourhood, its attractiveness and building capacity. By employing ML and geovisualisation techniques, the study revealed underlying patterns and uncovered hidden connections among various demographic groups that could otherwise only be inferred.

This study analysed and connected urban and socio-economic attributes while identified trends within different population groups, urban environments and specific time periods. Consequently, it provided a basis for optimising the process of selecting the appropriate combination of technological tools and data to produce fine-grained populations. This approach allowed the generation of precise estimates for the location and population density of areas resembling the dimensions of a typical large European city block shedding light on the past, present and future population distribution.

The following subsection emphasises the contributions of the presented papers in regards to the research questions. Papers A and B contributed partially and complementary to each of the first two questions, while Papers C and D contributed to the last research question. Limitations of this study, ethical considerations and suggestions for future work are presented in sections 6.2, 6.3 and 6.4.

6.1 Contributions to the research questions

RQ1: How can multi-output ML models improve the disaggregation of multiple population variables over classic dasymetric approaches and single-output models?

Papers A and B presented a spatial disaggregation approach using self-training and experiments over various regression models. The approach aimed to downscale counts of multiple interrelated variables to fine-grained grids in urban areas by leveraging ML models. By comparing the results with classic pychnophylactic and dasymetric interpolation methods as baselines and employing them as '*teacher models*', the papers examined whether ML models enhanced the quality of the disaggregated estimates and explored the mechanisms behind this enhancement.

Paper A presented experiments using ensemble learning (i.e. random forest, gradient boosting trees) in a single- and multi-output setup and Paper B extended the multi-output experimental setup to convolutional neural networks. A task specific loss function guided the training in combining the different variables in both cases.

These ML models could effectively handle the disaggregation of multiple interrelated variables, potentially correlated, where capturing their relationships was crucial. Regression models, known for predicting continuous variables, were well-suited for downscaling aggregated counts of populations to fine-grained grids. The self-training technique, a form of semi-supervised learning, enabled the model to iteratively improve its performance by generating training data from its own predictions.

The papers showed valuable methodological contributions in the field of spatial disaggregation modelling, especially for studies involving multiple sociodemographic variables. The approach offered a more convenient and efficient downscaling process. Instead of computing single variables separately, this more holistic approach handled interconnected variables, producing estimates for multiple population groups at once.

They also significantly outperformed traditional methods, enhancing accuracy and spatial variability. By breaking the smooth interpolation of commonly used methods, this approach achieved results with greater spatial heterogeneity that more closely resembled the actual population distribution. Especially, the experiments presented in Paper A and B showed that the use of more sophisticated ML models, such as deep neural networks (DNNs), increased the desired spatial variability and predictive accuracy of the disaggregated estimates. The modelling approach provided increased flexibility and eliminated the need to repeatedly train numerous single-output models. This saved computational resources and reduced the time required for subsequent analysis.

RQ2: How can different assessment methods assist the spatial evaluation of disaggregated estimates when ground truth data are available at target resolution?

Following the development of a ML model, its evaluation is a crucial step in assessing its predictive accuracy and overall performance. The MAE and MAPE are traditional metrics that offer quantitative insights into the model's predictive quality. However, assessing the spatial error and overall quality in geospatial modelling requires a combination of multiple, and potentially unconventional, methods to evaluate *where* the error occurs. Comprehensive evaluation techniques can lead to a deeper understanding of the models' capabilities. The spatial interpretation of errors is a key challenge to enhancing the generalisability and accuracy of spatial models, whereas understanding the spatial implications of errors is crucial to improving the model's performance in real-world scenarios.

Papers A and B presented a quality assessment of the downscaled results by combining multiple metrics and spatial evaluation methods. In contrast to the commonly used evaluation approach that involves intermediate aggregation levels, these studies benefited from obtaining access to population registers from the national statistics offices for the corresponding groups, aggregated at the target resolution of 100 m. The study showed a more precise and direct method that compared the model's predictions with the ground truth population distribution at the same resolution, achieving an accurate one-to-one evaluation of its quality. It is important to highlight that the ground truth data were not used in the training of the models to resemble situations where the corresponding data are not available.

In Paper A, the authors employed the MAE and the MAPE to compare the models, while they also visually inspected the produced estimates against the ground truth distribution. This visual inspection of the produced maps revealed differences in the variability of the results and distinguished those models that significantly outperformed others. Nevertheless, in certain situations where the models produced results with minimal differences, the pixel-level evaluation did not yield conclusive results. Moreover, the visual inspection was particularly challenging in densely populated areas because of the small size of the grid cells in large study areas.

To overcome these limitations and to facilitate the interpretation of the spatial error, in Paper B, the authors proposed a combination of alternative quality assessment methods, including the distribution of error by land use type and in groups of neighbouring cells, in addition to the traditional ones. Specifically, the use of the concept of individualised neighbourhoods proved very useful in balancing over- and underestimation errors in neighbouring cells. Instead of focusing on individual pixels, these visualisations reflected larger variations at zone level.

The combination of multiple methods and the integration of methods borrowed from the field of segregation analysis allowed a more comprehensive comparison of the differences of the models. Such an analysis helped to identify the most suitable model for specific tasks and shed light on the model's generalisability. Furthermore, different methods revealed different weaknesses and provided a more complete understanding of the model's capabilities, enabling a guide for addressing these limitations. They also gave unique perspectives on the ML model's performance and behaviour, uncovering previously unnoticed patterns, errors and areas for improvement.

For instance, evaluating the models through the relative error estimation (REE) across different land use categories provided a comprehensive overview of how well each model performed in different land use contexts, but also highlighted areas of specific land use where the models exhibited significantly large scale errors. Although the predictions of the two models examined appeared similar at the pixel level, the spatial error showed the differences at the zone level. A visual assessment of the spatial error revealed discrepancies in the extent and the magnitude of the error, while also more effectively highlighted the differences in the error distribution among the migrant groups.

Conducting such an analysis contributed to determine the robustness of the models under different scenarios and data distributions, and to understand how the model performed under diverse conditions, which is critical for real-world applications. It helped to tailor and refine the models to perform optimally under different environmental and land use conditions. It also highlighted research gaps and opened up new avenues for investigation and improvement, while also pointing or inspiring the direction of future research in ML development and evaluation.

Overall, exploring unconventional assessment methods for ML models can enrich the understanding of these models, especially for black-box algorithms, and lead to innovative approaches for evaluating and improving their performance. It can foster a deeper comprehension of the strengths and limitations of ML algorithms, facilitating more informed and effective development for spatial applications, such as spatial disaggregation.

RQ3: How can geovisualisation and ML techniques help to reveal patterns in the distribution of different migrant groups in relation to the native population and other socio-economic and urban characteristics?

A primary objective of this project was to explore the underlying patterns in migrant distributions and comprehend the factors that drive them, whether they act as push or pull mechanisms in particular regions of destination cities. This was a crucial step in collecting the appropriate data to train the disaggregation model, improve its performance and ensure accurate results. Papers C and D contributed to this research question, the former by proposing the RF VIMs to explore the factors influencing the residential locations of migrants, and the latter by mapping migration spatial statistics to analyse migration patterns.

A review of the existing literature on migrants' residential choices showed that most studies in this area follow qualitative methods based on surveys. Instead, Paper C suggested a data-driven approach to quantify the weight of importance of various factors identified in the review, including ethnic and demographic composition. Leveraging the capabilities of ML modelling, it analysed the spatial distribution of various migrant communities and uncovered the underlying drivers behind their residential locations, finding evidence that the ethnic composition of a neighbourhood was important to many minorities for the year under review, as revealed by their mutually high importance scores.

The most prominent contribution of this study was the identification of key drivers in modelling migrants' residential location and the quantification of their impact. By ranking the feature importance, the research provided valuable insights into the factors that played an important role in migrants' residential distributions in 2018. The used measure quantified the importance of different features on the model's predictions, allowing the comparison of the relative influence among ethnic minorities. Understanding these key factors and their weight of significance can lead to a better understanding of migration processes and inform future studies of human mobility.

The analysis also contributed to the field of migration studies and spatial analysis by advancing the state of the art in migrant residential location modelling through the use of sophisticated data-driven techniques such as the RF algorithm and VIMs. The approach demonstrated the suitability and effectiveness of these methods in capturing complex migration patterns in urban environments. The findings of the research validated previous qualitative studies related to migrant residential location and extended existing knowledge to a higher level of detail, examining migrants from specific CoOs. Such an approach opens up the potential to test existing theories through data-driven research and to develop new theories on the basis of these tests. For example, applying this method over multiple time periods could reveal whether spatial assimilation [93, 24] is occurring, among other potential patterns.

Furthermore, the research contributed to the field of XAI and the interpretability of the migrant residential location model. The study identified the most important features that make a ML model and its predictions more transparent, understandable and socially acceptable, while facilitating communication among policy makers, stakeholders and the academic community. Such communication can lead to practical implications for urban planning, migration management and policy development. Decision-makers can use the obtained knowledge to devise targeted interventions and policies that address migration-related challenges (e.g. parallel societies) and foster better integration of migrants into the urban fabric.

Following a geovisualisation-based approach, Paper D explored historical migration dynamics by generating meaningful choropleth, bivariate and proportional symbol maps. The research related two or more variables and effectively communicated connections between data characteristics, revealing differences in the distributions of different migrant, socio-economic and demographic groups, inner city mobility patterns and urban amenities. The paper's primary contribution lied in its methodological approach to mapping gridded data and in demonstrating the use of geovisualisation techniques and bivariate mapping as effective tools for studying migration patterns.

The paper advanced knowledge in the field of geography by providing insights into these patterns, historical trends and spatial relationships using advanced geovisualisation methods with an emphasis on aesthetics and effective communication. It contributed to cartography by showcasing creative ways of representing gridded migration data on maps. Moreover, the use of statistical calculations and bivariate mapping techniques contributed to the field of data science by demonstrating how visual representations can effectively communicate rich but complex data.

The paper enhanced the understanding of migration patterns in destination cities over different time periods, including factors that determined the distribution of migrants, such as the ethnic and social composition of an area and access to recreational, educational and cultural services. Finally, the visually compelling nature of the proposed geovisualisations not only provided insights into the complex dynamics of human migration and its wider implications, but also the means to communicate them to a wider audience, raise public awareness and promote informed discussions on migration-related issues.

6.2 Limitations

The present research has yielded interesting results and delved into a wide range of topics. However, various limitations were encountered, which are discussed in this section. These limitations refer to the case study areas, the disaggregation methodology and the high-resolution datasets employed.

Considerations on case study areas

Throughout the project, the uniqueness of each case study and the variations in their requirements and possibilities became apparent. The availability of the necessary datasets differed across cities in terms of time frame, scope and administrative levels. Some datasets dated back to the early 1990s, while others were more recent, starting from 2010. The geographical coverage also varied, with some cases extending over a single municipality and others covering multiple municipalities.

One of the primary challenges was to determine the definition of migrants and which classes of migrants were to be included in the analysis. Various sources define migrants by their CoO, CoB or current citizenship. CoB refers to the country in which a person is born. It is a static characteristic that does not change over time and is often used as a demographic indicator to analyse the composition of populations. CoO refers to the country from which a person or their ancestors migrated. It represents the original nationality of the individual and can change over generations as individuals or their descendants move to different countries or acquire new nationalities. It is a dynamic characteristic and is often used in migration studies to examine patterns of migration, integration and assimilation.

In our case, local authorities were requested to provide historical data based on either CoO or, if this information was not available, country of citizenship (CoC). This approach allowed capturing the presence of descendants, to understand the complex dynamics of migration and to examine past segregation patterns. However, this decision led to inconsistencies between the national and regional projections and the local distributions, since the former were estimated on the basis of CoB and the latter on the basis of CoO/CoC. Conversion matrices based on historical migration patterns and data were used to convert the regional projections from CoB to CoO, introducing another possible source of inaccuracy.

Another challenge was the large number of countries included in the analysis and the need to classify them appropriately with the perspective of conducting a comparative study between cities. Several approaches were employed and tested to achieve a homogeneous classification, including the M49 standard for countries and geographical regions commonly used by the United Nations Statistics Division [UNSD; 150]. However, due to the unique histories of the cities and their different patterns of attracting migrants from different regions, a classification tailored to each individual case study was ultimately adopted.

Disaggregation methodology and produced estimates

Considering the limitations associated with the disaggregation methodology, the experiments conducted for Paper A and B showed that the use of more sophisticated ML models, such as DNNs, increased the spatial variability and consequently the predictive accuracy of the disaggregated estimates. Though, the complexity of the configuration also increased. Many parameters had to be tuned and refined, increasing the challenges of implementation and reproducibility. Concerns were also raised about the interpretability of the model, especially of a black-box model, making the interpretation of results less straightforward than with ensemble modelling.

The findings of Paper B highlighted an important limitation of using DNNs in certain contexts. While they may have performed well when provided with detailed datasets of building features, their sensitivity to such data limits their implementation possibilities in real-world scenarios where high-resolution training datasets are not readily available in most countries. This limitation has implications for the generalisability of the model to different regions with different building types, the exacerbation of biases in the training data and the cost of data collection.

Paper A and B presented results that relied on initial weights deriving from the GHS-POP layer and highly depended on its quality. As discussed in Paper B, the predictions of the examined models encountered challenges in regions with a disproportionately high share of migrants. The size of migrant communities in these areas was underestimated due to the initial weights being distributed proportionally to the total population. This limitation affected the accuracy of the predictions in such regions.

Additionally, the proposed methodology did not yield meaningful estimations for very small population groups. Experiments were conducted in the case study of Kraków, focusing on population groups with sizes ranging from 100 to 800 persons. The predictions for the entire municipality resulted in gridded cells with values ranging from 0 to 1 person, which indicates an implausible outcome.

Despite the good results achieved for one reference year as presented in Papers A and B, discrepancies were noticed during the disaggregation of the statistical projections for multiple steps in the future – outputs produced as part of the FUME deliverables. These outputs could not correspond to realistic sequential distributions, as shown by the segregation analysis of the historical data and the disaggregated estimates for the future performed in the context of the final stages of FUME. These inconsistencies and deviations were both related to the use of the specific weights in guiding the disaggregation and the lack of guidance from the previous time step.

Because of these limitations, it is mentioned in section 5 that adjustments were made to improve the consistency of the results. Consistency was ensured by replacing the GHS-POP layer with the most recent distribution of the corresponding group, on the assumption that the already attractive areas are likely to remain attractive in the subsequent steps and that the existing population will not significantly move out of these areas. In addition, the disaggregation was adjusted to reflect the difference in the size of the population of each group between the periods studied, rather than their total size. In cases where the population would decrease, the decrease was assumed to be more likely for grid cells with a higher population, which may not accurately reflect reality.

Due to the large number of individual countries of origin involved, it was considered impractical to attempt to estimate predictions for each individual country. This was due to the small numbers of individuals originating from most countries and their absence in various regions, which would not yield meaningful results. Consequently, the study took a different approach by estimating the distribution of migrants based on RoOs instead of individual countries. This adaptation allowed for flexibility according to the possibilities and research interests of each case study, but resulted in less detailed local projections.

Running ML simulations can be computationally demanding and resource intensive, especially for the DL models. It requires a suitable environment with high performance hardware, such as graphics processing units (GPUs); therefore, a virtual environment was created using CLAAUDIA [66], which offers substantial advantages in terms of storage and computational capacity. These simulations could not be supported though within the DST environment, necessitating an approach that did not rely on historical data, which were strictly accessed in the DST environment. Even in the CLAAUDIA environment, the CNN experiments encountered memory issues when attempting to significantly increase the number of groups under examination.

Moreover, exploring DL solutions with a growing number of parameters led to an increase in the required experiments. The increased number of experiments and necessary computational power raised concerns about the environmental impact of ML applications, questioning its alignment with the concept of '*green AI*'. The responsible use of artificial intelligence for sustainability minimising its carbon footprint and energy consumption should be prioritised in the ML research agenda in the future.

Finally, it is important to recognise DL's vulnerability to deception in general. Neural networks can learn patterns effectively, but they can struggle to see the bigger picture. If the data lacked representative patterns, or if certain patterns went unnoticed, the model could not recognise what it had not learned. It was also unable to predict patterns that did not exist in the data (e.g. in the case of new influxes of migrants).

Considerations on high-resolution datasets

Limitations related to the availability and accessibility of the fine-resolution population datasets were uncovered in the presented papers and this dissertation. Access to such a detailed dataset was one of the main advantages of the present research, offering unique opportunities to investigate the distribution of various migrant groups. Though, the direct dependence on ground truth data at the resolution of 100 m introduced implications throughout the project for evaluating the disaggregated output in Paper A and B, for examining the importance of features/factors influencing the distribution of migrants in Paper C and for analysing the migration patterns with visualisation techniques in Paper D.

The closed nature of the population datasets at the specific resolution presented difficulties in pre-processing and cleaning. The datasets of the case studies were highly heterogeneous, lacked proper documentation and exhibited inconsistencies, requiring extensive communication with the local and FUME partners and time-consuming pre-processing and cleaning steps, resulting in occasional discrepancies. Additionally, the closed datasets limited collaboration with partners, that could provide valuable insights, and reproducibility from third parties, making verification and validation of the results challenging.

To address these difficulties, including the challenges posed by the small cell size as described in Paper B, an evaluation method was adopted that involved aggregating the error in zones of neighbouring cells. While this approach helped interpret the distribution of the error, it introduced issues at the local scale if zone sizes were not carefully selected. Additionally, the evaluation process resulted in the creation of numerous maps, which demanded considerable time for interpretation and analysis. In contrast, tables offered direct comparison metrics among the models and variables.

In the case of Paper D, the population and topographic data were aggregated to 1 km grid cells to achieve a visually appealing and comprehensive result, that would not be possible with the fine resolution of 100 m. A crucial aspect to bear in mind is that the produced grid cells were suitable for only a few cartographic visualisation methods (i.e. choropleth, bivariate, and proportional symbol maps). Challenges were encountered in finding suitable colour schemes to effectively communicate data relationships while adhering to cartographic conventions for representing natural phenomena.

6.3 Ethical considerations

Analysing the residential locations of various migrant groups or projecting their future distribution can have negative implications, especially if not handled with sensitivity and ethical considerations. Specifically, privacy concerns have been raised, even though the historical data were aggregated to grid cells. The spatial resolution, together with the details of the CoOs, may have resulted in the exposure of sensitive information about individuals or communities, leading to potential privacy violations, if the data were mishandled.

In today's world, where nationalistic sentiments are on the rise, population projections serve as powerful tools and come with inherent risks. These risks include the potential for governments to manipulate or tune these models to align with policies driven by xenophobia and discrimination. The projections for the future distributions of certain migrant groups can lead to stigmatisation and discrimination against them. When such projections are misused, they may perpetuate stereotypes or contribute to negative perceptions about particular communities, leading to social tensions and marginalisation. They may set the grounds for exacerbated segregation and clustering threatening cross-cultural interactions and the loss of cultural identities.

In addition, such information on the future residential locations may attract real estate speculations and gentrification procedures in areas where the concentration of migrants is higher. Rising property values could lead to the displacement of the original residents, including both migrant and native residents, who might no longer afford to live in those areas.

Different risks may be hidden when such information, either accurate or inaccurate, is used by local authorities and stakeholders for urban planning decisions. Specifically, resources might be unequally distributed with some areas receiving disproportionate attention and resources compared to others where diverse or less visible migrant populations are overlooked. Exacerbating existing disparities or excluding vulnerable populations from services and resources are additional risks that need to be considered.

By tuning models to various scenarios, governments can gain valuable insights into how their cities might develop under different conditions. It is also important to ensure that these projections are applied in a responsible and ethical manner, prioritising the well-being and inclusion of all communities. However, relying solely on such projections can lead to misguided urban planning decisions that do not reflect the actual needs and dynamics of the population. In such cases, additional factors should also be considered, as the projections might be poor in accuracy or outdated. The engagement and active participation of migrant communities in planning processes can promote the needs and preferences of diverse groups for effective planning built with trust and representation.

6.4 Future research and outlook

Research is an never-ending journey and this dissertation is no exception. While this dissertation presents valuable contributions and scientific outcomes, its research findings also pave the way for new challenges, research areas and innovations. At several points during the PhD project, different tracks could have been selected if not for time constraints or data unavailability. With this in mind, I would like to suggest potential tracks for future work in terms of refining and validating the proposed approaches, but also in terms of methodological alternatives and other interesting topics for further research.

Future research can focus on further refining and validating the multi-output disaggregation methodology. This can include exploring alternative supervised or unsupervised regression algorithms, incorporating additional ancillary variables and conducting extensive validation studies across various urban areas with diverse socio-demographic characteristics. Improving the generated estimates produced for specific migrant groups is a key concern. More reliable estimates could be achieved by integrating diverse data sources (e.g. remote sensing, social media data), the diaspora effect or other dynamic factors (e.g. economic trends, policy changes) into the training process or weighted inputs. Future studies can explore methods to integrate these factors into the spatial disaggregation approach to provide more comprehensive and accurate population estimates. Experiments with more time steps are also necessary for generalising the approach to historical data, with the exploration of time-specific datasets playing an important role.

In terms of uncertainty analysis, assessing and quantifying the uncertainty associated with the local estimates is crucial not only for the development of more generalisable and accurate spatial models, but also for informed decision making. Future studies can focus on systematic methods for evaluating and interpreting spatial errors when ground truth data are available at a fine-grained target resolution. These methods may include correlation analysis, error or anomaly detection and cross-validation. Residual analysis, which examines the correlation between the error and the ground truth, can provide valuable information for evaluating a ML model. It helps to determine whether the model is effectively capturing the underlying patterns, missing important features, or whether its performance varies across the data range. Object detection can be applied for error localisation, while anomaly detection can identify unusual patterns that fall outside certain thresholds, such as non-residential towers. Cross-validation can provide a more robust estimate of the model's performance, as certain areas are excluded from the modelling process for evaluation. The incorporation and analysis of model uncertainty in spatial disaggregation can provide policy makers and planners with more reliable and informative estimates.

The applicability of the methodologies of Paper C and D can also be explored in different contexts, including areas with different urban structures and regions with different demographic trends and urbanisation patterns (e.g. metropolitan regions in Asia and the Americas). Extending the studies to multiple time steps could provide evidence on when the concentration of one ethnic group becomes an attractive factor for another minority. It could also shed light on segregation trajectories involving the formation of ethnic enclaves or spatial assimilation. Its application to different regions and time periods may also contribute to the generalisability and transferability of the model, increasing confidence that the identified drivers remain consistent across different contexts. Sequence analysis could be a valuable adjunct to explore these aspects further, while geographically weighted regression offers opportunities to examine non-stationary variables and to model the local relationships among the variables.

As urban areas continue to evolve, long-term population projections and scenario analysis will become increasingly important. Future research can explore methods to explicitly generate long-term projections by working with time series of rasters. RNNs, LSTMs, gated recurrent units (GRUs) are designed to process sequential data, making them suitable for time series analysis, while combined with CNNs could improve the accuracy and granularity of local population projections. For example, convolutional long short-term memory networks (ConvLSTMs) have already been used to process time series of remote sensing images [81, 63, 131] and bear potential for applications on population projections.

During the final stages of the FUME project, time constraints limited the in-depth examination of the connection between the aggregated projections and the local distribution of migrants. While consistency between the projections and historical data was restored, the impact of different scenarios at the local level remained underexplored. Rather than changing the methodological approach with one of the ML alternatives mentioned above, a more systematic focus on quantifying various scenarios at the local level and investigating their implications is also suggested. This could involve developing just one scenario and conducting experiments to assess the impact of locally adapted scenarios, such as experiments with the same area for luxury versus social housing, or with varying building density coefficients. A sensitivity analysis of the impact of these dependent variables would be particularly interesting. This analysis can help to determine the robustness of the black-box model used and assess how the variables might affect the distribution of different groups. Such an exploration could provide valuable insights into the dynamics of the local distribution and reveal interesting variations in the findings.

Collaboration between researchers, policy makers, urban planners and data providers is essential to advance the field of local population estimates and projections. By promoting collaboration and sharing best practices, the field can benefit from collective knowledge and expertise, leading to more accurate and reliable results. Establishing standardised methodologies and benchmarks for spatial disaggregation and population projections can further enhance the consistency and comparability of research findings.

In order to promote the adoption of such tools by local authorities, it is essential to understand their specific needs and requirements. Further investigation into the needs of local authorities can provide practical insights and suggestions on how to make population estimations and projections more relevant and efficiently applicable in urban planning. Workshops and seminars can serve as platforms to engage with local authorities, stakeholders, academics and migrant communities to gain a deeper understanding of their perspectives and concerns.

Making population estimates and projections accessible to the public and local authorities can increase their usability and impact. Online tools and interactive mapping platforms can facilitate the dissemination of findings and allow stakeholders to interact with the data and better understand the implications. Last but not least, a deeper understanding of the interactions between different sectors can be achieved by integrating other models such as climate change, food/water demand, sustainability and emissions models into such platforms. Scenario-based assessments that explore possible futures can help to make informed decisions and develop proactive policies.

7

Conclusion

Conclusion

Cities are amazing places, attracting diverse people and cultures. They foster an environment of remarkable dynamism and vibrancy. In certain aspects, they are like foundries; dynamic environments where diverse elements come together and create something new and transformative. This process of synthesis, refinement and interaction leads to the emergence of new combinations that are greater than the sum of its parts.

Cities serve as crucibles where people from different backgrounds interact, exchange knowledge and contribute to the growth of culture, technology and society. This multicultural interaction enriches the cultural fabric of urban centres, stimulates intellectual discourse and fosters the open-mindedness that drives social progress. In this way, cities allow human diversity to unfold harmoniously, constantly redefining and rejuvenating the narrative of collective human achievement.

Nevertheless, cities exceed the complexity and dimensions of foundries, encompassing social, economic, political and cultural dimensions that go beyond the material processes of a foundry. Unlike castings, cities resist confinement and relentlessly challenge limitations, introducing an unpredictable element into their future trajectory. Their sustained well-being demands a strategic blueprint extending beyond conventional business plans, ensuring a secure, inclusive and sustainable place for their inhabitants.

Such complex processes require improved urban planning and decision-making. More effective planning, in turn, relies on better analysis of existing urban patterns and population dynamics, and on evidence-driven preparedness for unpredictable scenarios. Currently, there is a lack of established methodologies that can provide insights onto the factors influencing migrant settlement, impeding cities' ability to predict where future migrants might settle in. However, emerging technological advances, particularly ML, offer significant potential to address these challenges. Local societies and economies can be profoundly influenced by adopting interdisciplinary strategies that combine demographic research with geospatial and computational analysis, with a specific focus on migration and population projections.

In this way, cities have the opportunity to develop models and tools that yield valuable insights for decision-making and guidelines for inclusive urban development. These approaches cover various aspects, such as identifying optimal locations for migrant settlement, preventing segregation and ghettoisation, and considering the safety and attractiveness of neighbourhoods. This is especially valuable for the integration of new influxes of migrants or refugees, such as in the case of Ukrainians moving to EU countries.

In an attempt to overcome these challenges, this dissertation connected regional population projections and demographic estimates to local distributions of migrants in urban areas. It used the latest advances in ML and contributed to the field of GeoAI by extending existing models for socio-demographic applications.

With the first research question, this dissertation focused on exploring methods that facilitate the development of data-driven models for estimating the distribution of population groups on fine-grained grids. It contributed to the field of spatial disaggregation modelling by building upon existing tools, that were designed to provide more accurate estimates of migrant populations at high spatial granularity. By leveraging these tools, the study aimed to equip urban planning processes with improved foresight and decision-making capabilities.

The second research question highlighted the necessity of employing different assessment methods to spatially evaluate the disaggregated results. The use of different methods, including unconventional ones for the spatial interpretation of errors, was not only necessary to validate the accuracy of the modelling approach, but also enhanced the interpretation and reliability of the findings. It provided a comprehensive understanding of spatial patterns among different demographic groups, enabled validation and cross-verification of results, and provided insights into their spatial quality. Such insights were valuable for developing more accurate and generalisable spatially explicit models.

The final research question of the study contributed to the field of XAI and geovisualisation. XAI aims to make AI models more transparent and interpretable by providing explanations for their decisions. In this context, the third paper focused on explaining and validating the decisions made by the ML model in the study, based on existing knowledge. This ensured that the decisions made by the model could be understood and verified by human experts, while also quantifying relations that could not otherwise be measured in a data-driven way. The fourth paper explored migration and socio-economic dynamics through visual means and related them to urban features. The approach suggested the creation of informative and engaging visualisations for effectively identifying trends and patterns in gridded population data while enhancing the understanding and modelling of complex urban phenomena.

In essence, this PhD dissertation advanced the field of spatial disaggregation modelling while addressing the pressing need for accurate gridded population estimates, comprehensive spatial evaluation and interpretability of AI models. By integrating machine learning, spatial analysis and geovisualisation, it provided valuable tools and insights for a deeper understanding of migration dynamics in European destination cities. As part of the interdisciplinary project of FUME, this dissertation demonstrated a way to link aggregated population counts with local distributions to explore the potential socio-demographic consequences of migration based on different scenarios at the local level. It provided a starting point for exploring how cities might look like in the future under specific conditions and laid the foundation for further research and collaboration, promising more efficient, accurate and relevant population estimates for urban planning and policy-making in the future.

References

[1] Abel, G.J., 2010. Estimation of international migration flow tables in Europe. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 173 (4), 797–825. doi:10.1111/j.1467-985X.2009.00636.x.

[2] Abel, G.J., 2018. Non-zero trajectories for long-run net migration assumptions in global population projection models. *Demographic research*, 38 (1), 1635–1662. doi:10.4054/DemRes.2018.38.54.

[3] Abel, G.J. and Cohen, J.E., 2019. Bilateral international migration flow estimates for 200 countries. *Scientific data; Sci Data*, 6 (1), 82–13. doi:10.1038/s41597-019-0089-3.

[4] Abel, G.J. and Sander, N., 2014. Quantifying Global International Migration Flows. *Science*, 343 (6178), 1520–1522.

[5] Aitchison, J., 1981. A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13, 175–189.

[6] Aitchison, J., 2005. A concise guide to compositional data analysis. *In*: *Compositional Data Analysis Workshop*.

[7] Albert, A., *et al.*, 2018. Modeling urbanization patterns with generative adversarial networks. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018-July, 2095–2098. doi:10.1109/IGARSS.2018.8518032.

[8] Andersson, E.K., Lyngstad, T.H., and Sleutjes, B., 2018. Comparing patterns of segregation in north-western europe: A multiscalar approach. *European journal of population*, 34 (2), 151–168. https://www.jstor.org/stable/45178598.

[9] Azose, J.J. and Raftery, A.E., 2019. Estimation of emigration, return migration, and transit migration between all pairs of countries. *Proceedings of the National Academy of Sciences - PNAS; Proc Natl Acad Sci U S A*, 116 (1), 116–122. doi:10.1073/pnas.1722334116.

[10] Balk, D.L., *et al.*, 2006. Determining global population distribution: Methods, applications and data. *Advances in Parasitology*, 62, 119–156. https://doi.org/10.1016/S0065-308X(05)62004-0.

[11] Barbosa, H., *et al.*, 2018. Human mobility: Models and applications. *Physics Reports*, 734, 1–74. https://doi.org/10.1016/j.physrep.2018.01.001.

[12] Barry, E. and Sorensen, M.S., 2018. In Denmark, Harsh New Laws for Immigrant ‘Ghettos’. https://www.nytimes.com/2018/07/01/world/europe/denmark-immigrant-ghettos.html.

[13] Black, W.R., 1995. Spatial interaction modeling using artificial neural networks. *Journal of Transport Geography*, 3 (3), 159–166. doi:10.1016/0966-6923(95)00013-S.

[14] Booth, H., 2006. Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22 (3), 547–581. doi:10.1016/j.ijforecast.2006.04.001.

[15] Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32. doi:10.1023/A:1010933404324.

[16] Buettner, T. and Muenz, R., 2018. *Modeling alternative projections of international migration*. KNOMAD working paper 3.

[17] Buettner, T. and Muenz, R., 2020. Migration projections: The economic case. *KNOMAD Paper Series No 37*.

[18] Burch, T.K., 2018. The Cohort-Component Population Projection: A Strange Attractor for Demographers. *In*: *Model-based demography*. Springer, Cham, 135–151. http://link.springer.com/10.1007/978-3-319-65433-1.

[19] Carey, H.C., 1883. *Principles of social science*. Philadelphia. http://hdl.handle.net/2027/nyp.33433081993093.

[20] Carvalho, D.V., Pereira, E.M., and Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics (Basel)*, 8 (8), 832. doi:10.3390/electronics8080832.

[21] CIESIN. Center for International Earth Science Information Network - Columbia University, 2016. Gridded Population of the World, Version 4 (GPWv4): Administrative Unit Center Points with Population Estimates. Accessed on 08.08.2022, http://dx.doi.org/10.7927/H4F47M2C.

[22] CIESIN. Center for International Earth Science Information Network - Columbia University, 2018. Gridded population of the world, version 4 (gpwv4): Population density, revision 11. Accessed on 08.08.2022, https://doi.org/10.7927/H49C6VHW.

[23] CBS. Centraal Bureau voor de Statitiek, 2023. Kaart van 100 meter bij 100 meter met statistieken. Accessed on 21.06.2023, https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken.

[24] Chihaya, G.K., *et al.*, 2022. Trajectories of spatial assimilation or place stratification? a typology of residence and workplace histories of newly arrived migrants in sweden. *The International migration review*, 56 (2), 433–462. https://journals.sagepub.com/doi/full/10.1177/01979183211037314.

[25] Coenen, A., Verhaeghe, P., and de Putte, B.V., 2019. Ethnic residential segregation: A matter of ethnic minority household characteristics? *Population space and place*, 25 (7). https://onlinelibrary.wiley.com/doi/abs/10.1002/psp.2244.

[26] Couronné, R., Probst, P., and Boulesteix, A.L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19 (1), 270. `https://doi.org/10.1186/s12859-018-2264-5`.

[27] Cushing, B. and Poot, J., 2004. Crossing boundaries and borders: Regional science advances in migration modelling. *Papers in Regional Science*, 83 (1), 317–338. `doi:10.1007/s10110-003-0188-5`.

[28] DST. Danmarks Statistik, 2019. *Immigrants in Denmark 2019 [Indvandrere i Danmark 2019]*.

[29] DST. Danmarks Statistik, 2023. Documentation of statistics: The Population - Statistics Denmark. Accessed on 04.04.2023, `https://www.dst.dk/en/Statistik/dokumentation/documentationofstatistics/the-population`.

[30] Davis, H.C., 1995. Migration Models. *In*: *Demographic projection techniques for regions and smaller areas : A primer*. UBC Press, 60–76.

[31] de Abreu Araújo, I., Torres, R.H., and Neto, N.C.S., 2022. A review of framework for machine learning interpretability. *In*: D.D. Schmorrow and C.M. Fidopiastis, eds. *Augmented*, Cognition, Cham. Springer International Publishing, 261–272.

[32] degli Uberti, S., *et al.*, 2023. SENEGAL: Drivers and trajectories of migration to Europe. July. `https://doi.org/10.5281/zenodo.8223962`.

[33] Diop, L.E.N., Kessler, C., and Basheer, S., 2023. IRAQ: Drivers and Trajectories of Migration to Europe. July. `https://doi.org/10.5281/zenodo.8189420`.

[34] Dobroczek, G., Puzynkiewicz, J., and Chmielewska, I., 2017. A new wave of ukrainian migration to poland | obserwator finansowy: Ekonomia | gospodarka | polska | Świat. -01-19T00:00:00+00:00. `https://www.obserwatorfinansowy.pl/in-english/new-trends/a-new-wave-of-ukrainian-migration-to-poland/`.

[35] Dobson, J.E., *et al.*, 2000. LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, 66 (7), 849–857.

[36] Docquier, F., 2018. Long-term trends in international migration: Lessons from macroeconomic model. *Economics and business review*, 4 (1), 3–15. `doi:10.18559/ebr.2018.1.1`.

[37] Elío, J., *et al.*, 2022. Migration studies with a compositional data approach: A case study of population structure in the capital region of denmark. *In*: O. Gervasi, B. Murgante, S. Misra, A.M.A.C. Rocha and C. Garau, eds. *Computational Science and Its Applications – ICCSA 202*, 2 Workshops, Cham. Springer International Publishing, 576–593.

[38] Entzinger, H.B., Scholten, P., and Crul, M., 2019. *A tale of two cities. rotterdam, amsterdam and their immigrants.* 173–189.

[39] Eurostat, 2020. Population projections - population and demography - eurostat. https://ec.europa.eu/eurostat/web/population-demography/population-projections.

[40] Eurostat, 2023. Population projections at regional level. 12 April. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_projections_at_regional_level.

[41] Ewers, M.C. and Dicce, R., 2018. High-skilled migration and the attractiveness of cities. *In*: *High-skilled migration: Drivers and policies*. Oxford University Press, 176–194. doi:10.1093/oso/9780198815273.003.0009.

[42] Fallov, M.A. and Birk, R.H., 2022. The 'ghetto' strikes back: resisting welfare sanctions and stigmatizing categorizations in marginalized residential areas in denmark. *Nordic Social Work Research*, 12 (2), 217–228. https://doi.org/10.1080/2156857X.2021.1937289.

[43] Fergus, P. and Chalmers, C., 2022. *Applied deep learning : tools, techniques, and implementation*. Cham, Switzerland: Springer.

[44] Fibæk, C.S., *et al.*, 2022. A deep learning method for creating globally applicable population estimates from sentinel data. *Transactions in GIS*, 00, 1–29. https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12971.

[45] Fischer, M.M. and Reggiani, A., 2005. Spatial Interaction Models : From the Gravity to the Neural Network Approach. *In*: *Urban dynamics and growth: Advances in urban economics*. vol. 266. Bingley: Emerald Group Publishing Limited Copyright © 2005, Emerald Group Publishing Limited, 319–346. https://doi.org/10.1108/S0573-8555(2005)0000266012.

[46] Fotheringham, A., 2001. Spatial Interaction Models. *International Encyclopedia of the Social & Behavioral Sciences*, 14794–14800. doi:10.1016/b0-08-043076-7/02519-5.

[47] Freiesleben, A.M.v., 2016. Et Danmark af parallelsamfund: Segregering, ghettoisering og social sammenhængskraft: Parallelsamfundet i dansk diskurs 1968-2013–fra utopi til dystopi.

[48] FUME. Future Migration Scenarios for Europe, 2023. Drivers and trajectories of migration to europe - summary. Accessed on 26.09.2023, https://futuremigration.eu/wp-content/uploads/2023/07/FUME-6.1-summary-publication-final.pdf.

[49] Gao, J., 2017. *Downscaling global spatial population projections from 1/8-degree to 1-km grid cells*. https://opensky.ucar.edu/islandora/object/technotes:553.

[50] Gao, J. and O'Neill, B.C., 2019. Data-driven spatial modeling of global long-term urban land development: The SELECT model. *Environmental Modelling and Software*, 119 (July), 458–471. https://doi.org/10.1016/j.envsoft.2019.06.015.

[51] Gemeente Amsterdam, 2023. Maps Amsterdam. Accessed on 21.06.2023, `https://maps.amsterdam.nl/`.

[52] Georgati, M., *et al.*, Submitted. Multi-output spatial disaggregation of population data using gradient boosting and deep learning models. *Transactions in GIS*.

[53] Georgati, M. and Keßler, C., 2021. Spatially Explicit Population Projections: The case of Copenhagen, Denmark. *AGILE: GIScience Series*, 2, 1–6. `doi:10.5194/agile-giss-2-28-2021`.

[54] Georgati, M., *et al.*, 2022. Spatial Disaggregation of Population Subgroups Leveraging Self-Trained Multi-Output Gradient Boosting Regression Trees. *AGILE: GIScience Series*, 3, 1–14. `https://agile-giss.copernicus.org/articles/3/5/2022/`.

[55] Giannakouris, K., 2010. *Regional population projections europop2008: Most eu regions face older population profile in 2030.* Publications Office of the European Union. `https://ec.europa.eu/eurostat/documents/3433488/5564440/KS-SF-10-001-EN.PDF.pdf/d5b8bf54-6979-4834-998a-f7d1a61aa82d?t=1414692757000`.

[56] Goodfellow, I., Bengio, Y., and Courville, A., 2016. *Deep learning.* Cambridge, Massachusetts: The MIT Press.

[57] Greenwood, M.J. and Hunt, G.L., 2003. The early history of migration research. *International Regional Science Review*, 26 (1), 3–37. `doi:10.1177/0160017602238983`.

[58] Géron, A., 2018. *Hands-on machine learning with scikit-learn and tensorflow.* First edition, ninth release ed.

[59] Hansen, H.S., 2018. Meeting the Migration Challenges at Local Governance Level by Small Scale Population Projections. *In*: A. Kő and E. Francesconi, eds. *Electronic government and the information systems perspective.* Springer. `doi:10.1007/978-3-319-98349-3`.

[60] Hapfelmeier, A. and Ulm, K., 2013. A new variable selection approach using random forests. *Computational statistics data analysis*, 60, 50–69. `https://dx.doi.org/10.1016/j.csda.2012.09.020`.

[61] Hastie, T., Tibshirani, R., and Friedman, J., 2009. *The elements of statistical learning data mining, inference, and prediction, second edition.* 2nd ed. New York, NY: Springer New York. `doi:10.1007/978-0-387-84858-7`.

[62] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9 (8), 1735–1780. `https://www.bioinf.jku.at/publications/older/2604.pdf`.

[63] Hu, W.S., *et al.*, 2020. Spatial–spectral feature extraction via deep convlstm neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (6), 4237–4250.

[64] IOM. International Organization For Migration, 2011. International Organization For Migration. Accessed on 14.09.2020, `https://www.iom.int/countries/denmark`.

[65] IOM. International Organization for Migration, 2021. *World migration report 2022.* Accessed on 23.07.2023, `https://publications.iom.int/books/world-migration-report-2022`.

[66] IT Services, Aalborg University, 2022. CLAAUDIA - Research Data Services. `https://www.claaudia.aau.dk/`.

[67] Jacob Schewe, Rikani, A., and Kluge, L., 2020. *International migration model. FUME Deliverable 4.2.*

[68] Janowicz, K., *et al.*, 2019. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34 (4), 625–636. `https://doi.org/10.1080/13658816.2019.1684500`.

[69] Jones, B. and O'Neill, B.C., 2016. Spatially explicit global population scenarios consistent with the Shared Socioeconomic Pathways. *Environmental Research Letters*, 11 (8). `doi:10.1088/1748-9326/11/8/084003`.

[70] Jones, B. and O'Neill, B.C., 2013. Historically grounded spatial population projections for the continental United States. *Environmental Research Letters*, 8 (4). `doi:10.1088/1748-9326/8/4/044021`.

[71] KC, S., *et al.*, 2022. *National level population and migration projections. FUME Deliverable 4.3.*

[72] KC, S. and Lutz, W., 2017. The human core of the shared socioeconomic pathways: Population scenariosn by age, sex and level of education for all countries to 2100. *Elsevier Ltd Global Enviromental Change*, 42, 181–192.

[73] KC, S., Speringer, M., and Wurzer, M., 2017. Working Paper Population projection by age , sex , and educational attainment in rural and urban regions of 35 provinces of India , 2011-2101 : Technical report on projecting the regionally explicit socioeconomic heterogeneity in India.

[74] KC, S., *et al.*, 2018. Future population and human capital in heterogeneous India. *PNAS*, 115 (33), 8328–8333. `doi:10.1073/pnas.1722359115`.

[75] Keßler, C. and Marcotullio, P.J., 2017. A Geosimulation for the Future Spatial Distribution of the Global Population. *AGILE 2017, 9.-12. May, Wageningen, The Netherlands*, 1–4.

[76] Koch, J. and Leimbach, M., 2023. Update of Ssp GDP Projections: Capturing Recent Changes in National Accounting, PPP Conversion and Covid 19 Impacts. *Ecological Economics*, 206. `http://dx.doi.org/10.2139/ssrn.4011838`.

[77] Kuffer, M., *et al.*, 2022. The Missing Millions in Maps : Exploring Causes of Uncertainties in Global Gridded Population Datasets.

[78] Kupiszewska, D. and Kupiszewski, M., 2005. *A Revision of the Traditional Multiregional Model to Better Capture International Migration: the Multipoles Model and Its Application.*

[79] LeCun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning. *Nature (London)*, 521 (7553), 436–444. `https://www.ncbi.nlm.nih.gov/pubmed/26017442`.

[80] Lent, M.V., Fisher, W., and Mancuso, M., 2004. An explainable artificial intelligence system for small-unit tactical behavior. *In*: *Proceedings of the national conference on artificial intelligence.* Citeseer, 900–907.

[81] Li, H.C., *et al.*, 2020. A 3 clnn: Spatial, spectral and multiscale attention convlstm neural network for multisource remote sensing data classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (2), 747–761.

[82] Long, L. and Zeng, X., 2022. *Beginning deep learning with tensorflow : work with keras, mnist data sets, and advanced neural networks.* New York, New York: Apress L. P. `doi:10.1007/978-1-4842-7915-1`.

[83] Lundsteen, M., 2023. Displacing the other to unite the nation: The parallel society legislation in denmark. *European urban and regional studies*, 96977642311652.

[84] Lutz, W., 1995. Scenario Analysis in Population Projection. *International Institute for Applied Systems Analysis*, (June), v, 14.

[85] Lutz, W., Butz, W.P., and KC, S., 2014. *World population and human capital in the twenty-first century.* Oxford: Oxford University Press. `doi:10.1093/acprof:oso/9780198703167.001.0001`.

[86] Lutz, W. and Goujon, A., 2001. The World ’ s Changing Human Capital Stock : Multi-State Population Projections by Educational Attainment. *Population and Development Review*, 27 (2), 323–339.

[87] Lutz, W., *et al.*, 2018. Demographic and human capital scenarios for the 21st century: 2018 assessment for 201 countries. 29113.

[88] Lutz, W. and KC, S., 2010. Dimensions of global population projections : what do we know about future population trends and structures ? *Philosophical Transactions of The Royal Society B*, 365, 2779–2791. `doi:10.1098/rstb.2010.0133`.

[89] Maestri, G., 2019. The nomad, the squatter and the state: Roma racialization and spatial politics in italy. *International Journal of Urban and Regional Research*, 43 (5), 930–946.

[90] Malmberg, B., *et al.*, 2018. Residential segregation of european and non-european migrants in sweden: 1990-2012. *European Journal of Population*, 34 (2), 169–193. `https://www.jstor.org/stable/45178599`.

[91] Marwal, A. and Silva, E.A., 2023. Exploring residential built-up form typologies in delhi: a grid-based clustering approach towards sustainable urbanisation. *npj Urban Sustainability*, 3 (1), 40.

[92] Massey, D.S. and Denton, N.A., 1988. The dimensions of residential segregation. *Social forces; Social Forces*, 67 (2), 281–315. `doi:10.1093/sf/67.2.281`.

[93] Massey, D.S. and Mullan, B.P., 1984. Processes of hispanic and black spatial assimilation. *The American journal of sociology*, 89 (4), 836–873. `doi:10.1086/227946`.

[94] McKee, J.J., *et al.*, 2015. Locally adaptive, spatially explicit projection of US population for 2030 and 2050. *Proceedings of the National Academy of Sciences of the United States of America*, 112 (5), 1344–1349. `doi:10.1073/pnas.1405713112`.

[95] Ministry of Immigration and Integration, 2019. *International migration- Denmark.*

[96] Molnar, C., 2022. *Interpretable machine learning.* 2nd ed. `https://christophm.github.io/interpretable-ml-book`.

[97] Monteiro, *et al.*, 2019. Spatial Disaggregation of Historical Census Data Leveraging Multiple Sources of Ancillary Information. *ISPRS International Journal of Geo-Information*, 8 (8), 327. `doi:10.3390/ijgi8080327`.

[98] Monteiro, J., *et al.*, 2021. Geospatial data disaggregation through self-trained encoder–decoder convolutional models. *ISPRS International Journal of Geo-Information*, 10 (9), 1–28.

[99] Monteiro, J., Martins, B., and Pires, J.M., 2018. A hybrid approach for the spatial disaggregation of socio-economic indicators. *International Journal of Data Science and Analytics*, 5 (2-3), 189–211. `doi:10.1007/s41060-017-0080-z`.

[100] Moss, R., *et al.*, 2008. *Towards New Scenarios for Analysis of Emissions, Climate Change, Impacts and Response Strategies.* No. December 2014. `http://www.osti.gov/energycitations/product.biblio.jsp?osti_id=940991`.

[101] Musterd, S., 2005. Social and ethnic segregation in europe: Levels, causes, and effects. *Journal of urban affairs*, 27 (3), 331–348. `doi:10.1111/j.0735-2166.2005.00239.x`.

[102] Nakicenovic, N., *et al.*, 2000. *Emissions Scenarios.* Cambridge: Intergovernmental Panel on Climate Change.

[103] NGR, 2023. Nationaal georegister. Accessed on 21.06.2023, `https://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/home`.

[104] Oak Ridge National Laboratory, 2021. Ornl landscan viewer. `https://landscan.ornl.gov/`.

[105] Olsson, G., 1970. Explanation, prediction and meaning variance: an assessment of distance interaction models. *Economic Geography*, 46, 223–233.

[106] O'Neill, B.C. and Balk, D., 2001. World population futures. *Population Bulletin*, 56 (3), 3–39.

[107] O'Neill, B.C., *et al.*, 2001. A Guide to Global Population Projections. *Demographic Research*, 4, 203–288. `doi:10.4054/demres.2001.4.8`.

[108] O'Neill, B.C., *et al.*, 2014. A new scenario framework for climate change research: The concept of shared socioeconomic pathways. *Climatic Change*, 122 (3), 387–400.

[109] Oomen, B., Baumgärtel, M., and Durmus, E., 2018. Transnational City Networks and Migration Policy Report by Cities of Refuge research. (December).

[110] Open Society Foundations, 2021. Denmark's Plan to Rebrand its Racist 'Ghetto Package' Will Cause More Housing Evictions - Open Society Justice Initiative. `https://www.justiceinitiative.org/newsroom/denmarks-plan-to-rebrand-its-racist-ghetto-package-will-cause-more-/housing-evictions`.

[111] Openshaw, S., 1979. A million or so correlated coefficients: three experiment on the modifiable areal unit problem. *Statistical applications in the spatial sciences.*

[112] Openshaw, S., 1984. The modifiable areal unit problem. *Concepts and techniques in modern geography.*

[113] Pellenbart, P.H. and Steen, P.J.M.V., 2001. Making space, sharing space: The new memorandum on spatial planning in the netherlands. *Tijdschrift voor economische en sociale geografie*, 92 (4), 503–512. `doi:10.1111/1467-9663.00175`.

[114] Philipov, D. and Rogers, A., 1980. Multistate Population Projections.

[115] Pomente, A. and Aleandri, D., 2017. Convolutional expectation maximization for population Estimation. *CEUR Workshop Proceedings*, 1866 (1).

[116] Pozniak, K., 2013. Generations of memory in the 'model socialist town' of nowa huta, poland. *Focaal*, 2013 (66), 58–68. `doi:10.3167/fcl.2013.660106`.

[117] Ravenstein, E.G., 1885. The Laws of Migration. *Journal of the Statistical Society of London*, 48 (2), 167–235.

[118] Rees, P., *et al.*, 2012. European Regional Populations : Current Trends , Future Pathways , and Policy Options. *European Journal of Population*, 28, 385–416. `doi:10.1007/s10680-012-9268-z`.

[119] Rich, D.C., 1980. Potential models in human geography. *CATMOG (Concepts & Techniques in Modern Geography)*, 26.

[120] Rikani, A. and Schewe, J., 2021. Global bilateral migration projections accounting for diasporas, transit and return flows, and poverty constraints. *Demographic research*, 45, 87–140. doi:10.4054/DemRes.2021.45.4.

[121] Robinson, C. and Dilkina, B., 2018. A machine learning approach to modeling human migration. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*. doi:10.1145/3209811.3209868.

[122] Robinson, C., Hohman, F., and Dilkina, B., 2017. A deep learning approach for population estimation from satellite imagery. *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, GeoHumanities 2017*, 47–54. doi:10.1145/3149858.3149863.

[123] Rogers, A., 1980. Introduction to multistate mathematical demography. *Environment and Planning A*, 12, 489–498.

[124] Roseveare, D. and Jorgensen, M., 2004. Migration and Integration of Immigrants in Denmark. http://dx.doi.org/10.1787/284832633602.

[125] Samir, K.C. and Lutz, W., 2017. The human core of the shared socioeconomic pathways: Population scenarios by age, sex and level of education for all countries to 2100. *Global Environmental Change*, 42, 181–192. doi:10.1016/j.gloenvcha.2014.06.004.

[126] Sapena, M., *et al.*, 2022. Empiric recommendations for population disaggregation under different data scenarios. *PLoS ONE*, 17 (9 Septamber), 1–29.

[127] Sassen, S., 2002. Global Cities and Disaporic Networks: Microsites in Global Civil Society. *Global Civil Society 2002*, 217–238.

[128] Schiavina, M., Freire, S., and MacManus, K., 2022. Ghs-pop r2022a - ghs population grid multitemporal (1975-2030). Accessed on 08.08.2022, http://data.europa.eu/89h/d6d86a90-4351-4508-99c1-cb074b022c4a.

[129] Schiavina, M., Freire, S., and MacManus, K., 2019. GHS-POP R2019A - GHS population grid multitemporal (1975, 1990, 2000, 2015). *European Commission, Joint Research Centre.* doi:10.2905/0C6B9751-A71F-4062-830B-43C9F432370F.

[130] Schug, F., *et al.*, 2021. Gridded population mapping for Germany based on building density, height and type from Earth Observation data using census disaggregation and bottom-up estimates. *PLoS ONE*, 16 (3 March), 1–23. http://dx.doi.org/10.1371/journal.pone.0249044.

[131] Singh, U., Gupta, P., and Shukla, M., 2022. Activity detection and counting people using mask-rcnn with bidirectional convlstm. *Journal of Intelligent Fuzzy Systems*, 43 (5), 6505–6520.

[132] Skaarup Larsen, H., Bach-Sørensen, N., and Breilev Lindgreen, T., 2018. *Projecting Spatial Population Distribution Using a Convolutional Neural Network*. Thesis (Masters). Aalborg University.

[133] Sleutjes, B., de Valk, H.A., and Ooijevaar, J., 2018. The Measurement of Ethnic Segregation in the Netherlands: Differences Between Administrative and Individualized Neighbourhoods. *European Journal of Population*, 34 (2), 195–224. `https://doi.org/10.1007/s10680-018-9479-z`.

[134] Smith, S.K., Tayman, J., and Swanson, D.A., 2013. *A Practitioner's Guide to State and Local Population Projections*.

[135] Sobczak-Szelc, K., Pędziwiatr, K., and Boubakri, H., 2023. TUNISIA: Drivers and trajectories of migration to Europe. July. `https://doi.org/10.5281/zenodo.8189334`.

[136] Soboleva, O., *et al.*, 2023. UKRAINE: Drivers and Trajectories of Migration to Europe. July. `https://doi.org/10.5281/zenodo.8189593`.

[137] CBS. Statistics Netherlands, 2022. New classification of population by origin. Accessed on 21.06.2023, `https://www.cbs.nl/en-gb/longread/statistische-trends/2022/new-classification-of-population-by-origin`.

[138] Stonawski, M., *et al.*, 2023. *Report on migration and residential segregation in European cities – now and future – analysis and projections of immigration and spatial distribution of immigrants. FUME Project Deliverable 6.4.*

[139] Stonawski, M., *et al.*, 2022. Investigating neighbourhood concentration of immigrants in poland: explorative evidence from kraków. *Bulletin of Geography. Socio-economic Series*, (56), 143–159. `https://apcz.umk.pl/BGSS/article/view/38008`.

[140] Storper, M. and Scott, A.J., 2009. Rethinking human capital, creativity and urban growth. *Journal of Economic Geography*, 9 (2), 147–167. `doi:10.1093/jeg/lbn052`.

[141] Stouffer, S.A., 1940. Intervening Opportunities: A Theory Relating Mobility and Distance. *American Sociological Review*, 5 (6), 845–867. `doi:10.1097/EDE.0b013e3181`.

[142] Striessnig, E., *et al.*, 2019. Empirically based spatial projections of US population age structure consistent with the shared socioeconomic pathways. *Environmental Research Letters*, 14 (11). `doi:10.1088/1748-9326/ab4a3a`.

[143] Swanson, D.A., *et al.*, 2009. Forecasting the population of census tracts by age and sex: An example of the hamilton-perry method in action.

[144] Sánchez Gassen, N., 2023. Improving the evidence for European migration policy making. June. `https://doi.org/10.5281/zenodo.8013345`.

[145] The Guardian, 2020. How Denmark's 'ghetto list' is ripping apart migrant communities | Migration | The Guardian. mar. Accessed on 20.07.2023, https://www.theguardian.com/world/2020/mar/11/how-denmarks-ghetto-list-is-ripping-apart-migrant-communities.

[146] Toth, F.L., Cao, G.Y., and Hizsnyik, E., 2003. *Regional population projections for China.* IIASA.

[147] UN DESA. United Nations, Department of Economic and Social Affairs, Population Division, 2019. *International Migration 2019: Highlights.* Accessed on 08.08.2023, https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2019_internationalmigration_highlights.pdf.

[148] UN DESA. United Nations, Department of Economic and Social Affairs, Population Division, 2022. World population prospects - population division - united nations. Accessed on 08.06.2023, https://population.un.org/wpp/.

[149] UN DESA. United Nations, Department of Economic and Social Affairs, Population Division, 2022. *World population prospects 2022: Summary of results.* Accessed on 08.08.2023, https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf.

[150] UNSD. United Nations, Department of Economic and Social Affairs, Statistics Division, 2023. UNSD — Methodology. Accessed on 08.08.2023, https://unstats.un.org/unsd/methodology/m49/.

[151] OHCHR. United Nations, Human Rights, Office of the High Commissioner, 2020. UN human rights experts urge Denmark to halt contentious sale of 'ghetto' buildings. Accessed on 01.09.2021, https://www.ohchr.org/en/press-releases/2020/10/un-human-rights-experts-urge-denmark-halt-contentious-sale-ghetto-/buildings.

[152] UN-Habitat. United Nations Human Settlements Programme, 2022. *World cities report 2022: Envisaging the future of cities.* United Nations.

[153] van Wissen, L., 2022. *Estimate Regional Migration and Population Scenarios at NUTS2 Level for DK, PL, IT and NL. FUME Deliverable 4.4.*

[154] Wang, D., 2015. Activating cross-border brokerage: Interorganizational knowledge transfer through skilled return migration. *Administrative Science Quarterly*, 60 (1), 133–176. doi:10.1177/0001839214551943.

[155] Wilson, T. and Rees, P., 2005. Recent developments in population projection methodology: A review. *Population, Space and Place*, 11 (5), 337–360. doi:10.1002/psp.389.

[156] World Bank, 2023. Urban development. Apr 03,. https://www.worldbank.org/en/topic/urbandevelopment/overview.

[157] World Economic Forum, 2017. Migration and Its Impact on Cities. *Journal of World Economic Forum*, 7 (October), 172. `http://www3.weforum.org/docs/Migration_Impact_Cities_report_2017_low.pdf`.

[158] WorldPop, 2022. WorldPop Data. Accessed on 08.08.2022, `https://www.worldpop.org/datacatalog/`.

[159] Yildiz, D., Wiśniowski, A., and Schewe, J., 2020. *Set of FUME Migration Scenario Narratives. FUME Deliverable 3.4.*

[160] Zipf, G.K., 1946. The P1 P2 / D Hypothesis : On the Intercity Movement of Persons. *American Sociological Review*, 11 (6), 677–686. `http://www.jstor.org/stable/2087063`.

[161] Zoraghein, H. and O'Neill, B.C., 2020. U.S. state-level projections of the spatial distribution of population consistent with shared socioeconomic pathways. *Sustainability (Switzerland)*, 12 (8). `doi:10.3390/SU12083374`.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-598-3

AALBORG UNIVERSITY PRESS